Analysis of Heart Disease Dataset

Justin Kim (yundong2@illinois.edu)

5/12/2021

Contents

Abstract

The goal of this analysis is test tools that can be used to screen for heart disease using methods learned in class. Machine learning methods such as KNN, decision trees, and random forest methods were used to build model and predict heart disease in patients. We will measure the accuracy of the models and decide which model to use when predicting the heart disease dataset.

Introduction

Heart disease is the leading cause of death in the United States. Heart disease refers to several types of heart conditions, but the most common case of heart disease is coronary artery disease. Coronary artery disease is the narrowing/blockage of major blood vessels supplying blood to the heart, possibly leading to heart failure. Through this analysis, we will build and pick the best possible model to predict heart disease in patients. ***

Methods

In order to model the dataset which character variables, we do some data cleaning such as factoring the character variables in order for it to work properly with machine learning methods. We also split the model into a training, test, validation, and estimation dataset

Data

The heart disease data set is from the UC Irvine Machine Learning Repository. This dataset was created by Hungarian Institute of Cardiology. * Budapest: Andras Janosi, M.D. * University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. * University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. * V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Below are the variables that will be used in the modeling of the dataset.

- age - the age of the patient
- sex - the gender of the patient
- cp - the type of chest pain experienced by the individual
- trestbps - the resting blood pressure
- chol - serum cholestoral
- fbs - fasting blood sugar
- restecg - resting electrocardiographic results
- thalach - maximum heart rate achieved
- exang - exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment
- num - number of major heart vessels with greater than 50% diameter narrowing
- thal: displays the thalassemia
- location - location of the heart disease

Results

```
## CART
##
## 240 samples
##  14 predictor
##   5 classes: 'v0', 'v1', 'v2', 'v3', 'v4'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 191, 192, 193, 192, 192
## Resampling results across tuning parameters:
##
##   cp           Accuracy   Kappa
##   0.000000000  0.5378528  0.24614471
##   0.007070707  0.5335975  0.22994824
##   0.014141414  0.5419308  0.23640087
##   0.021212121  0.5460124  0.23742161
##   0.028282828  0.5292571  0.17239070
##   0.035353535  0.5335124  0.17739997
##   0.042424242  0.5293458  0.17781130
##   0.049494949  0.5333351  0.18093253
##   0.056565657  0.5124132  0.09291772
##   0.063636364  0.5247431  0.09700499
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.02121212.
```

```
## k-Nearest Neighbors
##
## 240 samples
##  14 predictor
##   5 classes: 'v0', 'v1', 'v2', 'v3', 'v4'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 191, 192, 193, 192, 192
## Resampling results across tuning parameters:
##
##   k    Accuracy   Kappa
##    5   0.4755156  0.087533762
##    7   0.4963562  0.083208957
##    9   0.5336011  0.135439191
##   11   0.5086861  0.076559734
##   13   0.5253564  0.086638333
##   15   0.5463598  0.121683068
##   17   0.5336898  0.074016478
##   19   0.5547004  0.113617649
##   21   0.5669417  0.122263393
##   23   0.5628600  0.099087061
##   25   0.5585197  0.085002790
##   27   0.5375941  0.033434049
##   29   0.5374204  0.027983971
##   31   0.5375941  0.012666090
##   33   0.5500090  0.043639862
##   35   0.5417607  0.020896301
##   37   0.5417607  0.024517698
##   39   0.5500977  0.039103352
##   41   0.5460161  0.026962956
##   43   0.5460161  0.020015669
##   45   0.5417607  0.006523605
##   47   0.5417607  0.006523605
##   49   0.5417607  0.002972028
##   51   0.5417607  0.002972028
##   53   0.5417607  0.000000000
##   55   0.5417607  0.000000000
##   57   0.5417607  0.000000000
##   59   0.5417607  0.000000000
##   61   0.5417607  0.000000000
##   63   0.5417607  0.000000000
##   65   0.5417607  0.000000000
##   67   0.5417607  0.000000000
##   69   0.5417607  0.000000000
##   71   0.5417607  0.000000000
##   73   0.5417607  0.000000000
##   75   0.5417607  0.000000000
##   77   0.5417607  0.000000000
##   79   0.5417607  0.000000000
##   81   0.5417607  0.000000000
##   83   0.5417607  0.000000000
##   85   0.5417607  0.000000000
##   87   0.5417607  0.000000000
```

```
##    89  0.5417607  0.000000000
##    91  0.5417607  0.000000000
##    93  0.5417607  0.000000000
##    95  0.5417607  0.000000000
##    97  0.5417607  0.000000000
##    99  0.5417607  0.000000000
##   101  0.5417607  0.000000000
##   103  0.5417607  0.000000000
##   105  0.5417607  0.000000000
##   107  0.5417607  0.000000000
##   109  0.5417607  0.000000000
##   111  0.5417607  0.000000000
##   113  0.5417607  0.000000000
##   115  0.5417607  0.000000000
##   117  0.5417607  0.000000000
##   119  0.5417607  0.000000000
##   121  0.5417607  0.000000000
##   123  0.5417607  0.000000000
##   125  0.5417607  0.000000000
##   127  0.5417607  0.000000000
##   129  0.5417607  0.000000000
##   131  0.5417607  0.000000000
##   133  0.5417607  0.000000000
##   135  0.5417607  0.000000000
##   137  0.5417607  0.000000000
##   139  0.5417607  0.000000000
##   141  0.5417607  0.000000000
##   143  0.5417607  0.000000000
##   145  0.5417607  0.000000000
##   147  0.5417607  0.000000000
##   149  0.5417607  0.000000000
##   151  0.5417607  0.000000000
##   153  0.5417607  0.000000000
##   155  0.5417607  0.000000000
##   157  0.5417607  0.000000000
##   159  0.5417607  0.000000000
##   161  0.5417607  0.000000000
##   163  0.5417607  0.000000000
##   165  0.5417607  0.000000000
##   167  0.5417607  0.000000000
##   169  0.5417607  0.000000000
##   171  0.5417607  0.000000000
##   173  0.5417607  0.000000000
##   175  0.5417607  0.000000000
##   177  0.5417607  0.000000000
##   179  0.5417607  0.000000000
##   181  0.5417607  0.000000000
##   183  0.5417607  0.000000000
##   185  0.5417607  0.000000000
##   187  0.5417607  0.000000000
##   189  0.5417607  0.000000000
##   191  0.5417607  0.000000000
##   193  0.5417607  0.000000000
##   195  0.5417607  0.000000000
```

```
##   197  0.5417607  0.000000000
##   199  0.5417607  0.000000000
##   201  0.5417607  0.000000000
##   203  0.5417607  0.000000000
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 21.
```

```
## Random Forest
##
## 240 samples
##  14 predictor
##   5 classes: 'v0', 'v1', 'v2', 'v3', 'v4'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 191, 192, 193, 192, 192
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.5955873  0.2603888
##   10    0.5836047  0.3020662
##   19    0.5750127  0.2963099
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

After modeling, we see that knn method had the lowest accuracy of 0.5546, followed by decision tree at 0.5628, and random forest at 0.5919. From this, we choose the model with the highest accuracy, random forest.

Discussion

Because random forest had the best accuracy, we choose random forest method when predicting heart disease.
