STAT 3675 Final Project: Predicting People at Risk for Heart Attacks

Pavan Adapa

4/21/2021

Introduction

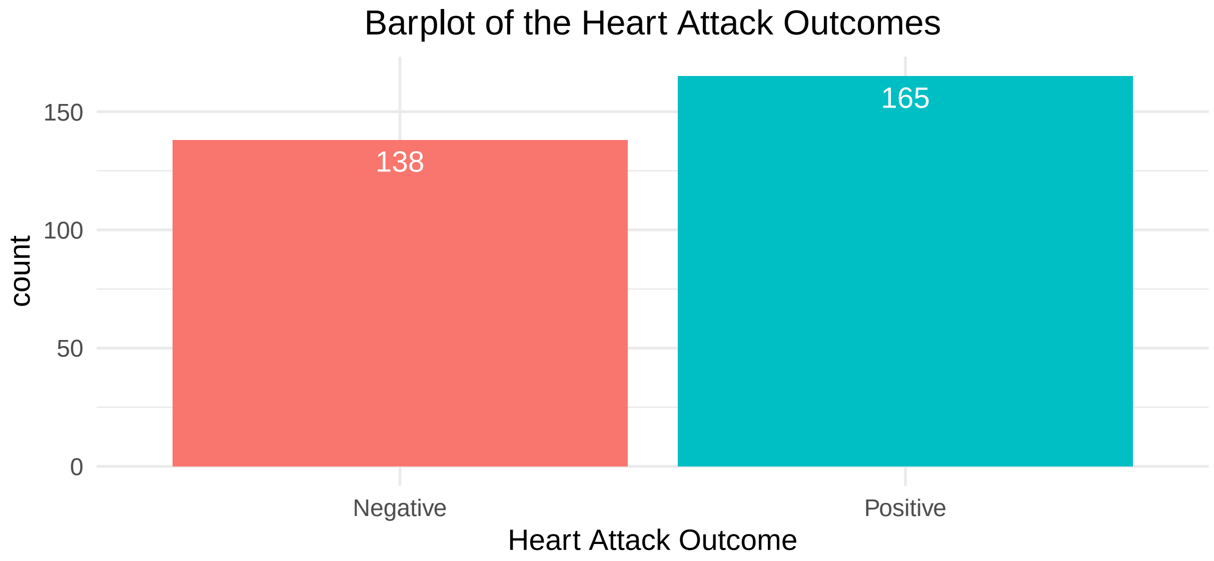
Healthcare is not only reactionary and it also entails preventative health. Preventative health is necessary to ensure longer lifespans and higher-quality lives. Millions of people die from preventable causes every year and increasing preventative care can end more needless tragedies. Classifications can be a very powerful tool for diagnosis as they can aid doctors where their biases might fail them otherwise. This project will attempt to predict people at risk for heart attacks by using several different classification methods and then pick the best model. The methods used will be logistic regression, classical decision trees, conditional inference trees, and random forest. The dataset was acquired from the University of Irvine Machine Learning Repository.

Elementary Data Anaylsis

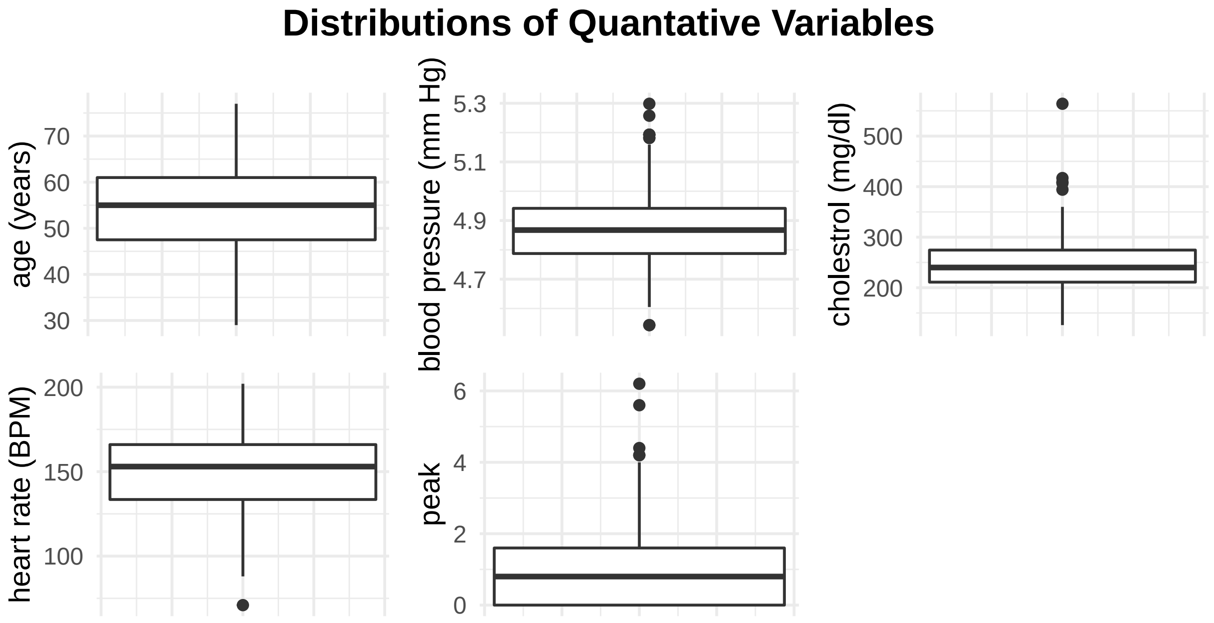
The dataset has 303 observations and of that data, approximately 80% will be the training data (240 observations). Therefore the remaining data (~20% or 63 observations) will the testing dataset. The dataset has 14 variables, and thus there are 13 independent variables. There are 7 categorical variables in the dataset of which two ordinal variables. The independent variables are: 1. Age: age in years
2. Sex: sex (1 = male; 0 = female)
3. Cp: chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic) 4. Trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. Chol: serum cholesterol in mg/dl
6. Fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. Restecg: resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy) 8. Thalach: maximum heart rate achieved
9. Exang: exercise induced angina (1 = yes; 0 = no)
10. Oldpeak: ST depression induced by exercise relative to rest
11. Slope: the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
12. Ca: number of major vessels (0-3) colored by fluoroscopy
13. Thal: Thalassemia (normal = 1; fixed defect = 2; reversible defect = 3) The independent variables is: Output: diagnosis of a heart attack (1 = yes, 0 = no)
Below is the summary of the data.

```
##       age        sex     cp          trtbps           chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
##  restecg    thalachh     exng       oldpeak     slp          caa
##  0:147   Min.   : 71.0   0:204   Min.   :0.00   0: 21   Min.   :0.0000
##  1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00   1:140   1st Qu.:0.0000
##  2:  4   Median :153.0           Median :0.80   2:142   Median :0.0000
##          Mean   :149.6           Mean   :1.04           Mean   :0.7294
##          3rd Qu.:166.0           3rd Qu.:1.60           3rd Qu.:1.0000
##          Max.   :202.0           Max.   :6.20           Max.   :4.0000
##  thall   output
##  0:  2   0:138
##  1: 18   1:165
##  2:166
##  3:117
##
##
```

1. Graphs Exploring Data



Interestingly the data has more occurrences of people having a heart attack than not.



While there are some outliers in the some of the distributions, none of the outliers have to be removed as most classification methods, including the ones used in this paper, doesn't require parametric distributions and thus are robust to outliers.

Logistic Regression

The first method of classification will be the logistic Regression. Below is the the original logistic model where no coefficients are dropped:

```
##
## Call:  glm(formula = formulaf, family = binomial, data = trainingdata)
##
## Coefficients:
## (Intercept)          age         sex1         cp.L         cp.Q         cp.C
##   4.4559507    0.0005028   -1.9286548    2.0539640   -0.3349036   -0.0326400
##      trtbps         chol         fbs1    restecg.L    restecg.Q     thalachh
##  -0.0218665   -0.0051756    0.2026866   -0.4053217   -0.5027862    0.0163724
##       exng1      oldpeak        slp.L        slp.Q          caa      thall.L
##  -0.3109810   -0.3855567    0.2897445    0.8830777   -0.7168357    0.2153682
##     thall.Q      thall.C
##  -1.9508809    0.9269859
##
## Degrees of Freedom: 239 Total (i.e. Null);  220 Residual
## Null Deviance:       331.9
## Residual Deviance: 166.9     AIC: 206.9
```

In the next few sections, all types of stepwise regressions will be performed to find the best model.

1. Backward Elimination

```
##
## Call:  glm(formula = factor(output) ~ sex + cp + trtbps + thalachh +
##     oldpeak + slp + caa + thall, family = binomial, data = trainingdata)
##
## Coefficients:
## (Intercept)         sex1         cp.L         cp.Q         cp.C       trtbps
##     3.55136     -1.79970      2.15369     -0.38044     -0.06896     -0.02343
##    thalachh      oldpeak        slp.L        slp.Q          caa      thall.L
##     0.01705     -0.43681      0.28953      0.92982     -0.70652      0.02574
##     thall.Q      thall.C
##    -1.94784      0.96015
##
## Degrees of Freedom: 239 Total (i.e. Null);  226 Residual
## Null Deviance:       331.9
## Residual Deviance: 170.1     AIC: 198.1
```

Backward elimination got rid of 5 variables. The variables removed where age, chol, fbs, restingecg, exng. Interestingly this model doesn't consider age as an important predictor for heart attacks. However, the AIC dropped form 205.3 to 196.5.

2. Forward Selection

```
##
## Call:  glm(formula = factor(output) ~ age + sex + cp + trtbps + chol +
##     fbs + restecg + thalachh + exng + oldpeak + slp + caa + thall,
##     family = binomial, data = trainingdata)
##
## Coefficients:
```

```
## (Intercept)          age         sex1         cp.L         cp.Q         cp.C
##   4.4559507    0.0005028   -1.9286548    2.0539640   -0.3349036   -0.0326400
##      trtbps         chol         fbs1    restecg.L    restecg.Q     thalachh
##  -0.0218665   -0.0051756    0.2026866   -0.4053217   -0.5027862    0.0163724
##       exng1      oldpeak        slp.L        slp.Q          caa      thall.L
##  -0.3109810   -0.3855567    0.2897445    0.8830777   -0.7168357    0.2153682
##     thall.Q      thall.C
##  -1.9508809    0.9269859
##
## Degrees of Freedom: 239 Total (i.e. Null);  220 Residual
## Null Deviance:       331.9
## Residual Deviance: 166.9     AIC: 206.9
```

Forward selection did not get rid of any variables and thus it the same as the orginal model.

3. Bidirectional Elimination

```
##
## Call:  glm(formula = factor(output) ~ sex + cp + trtbps + thalachh +
##     oldpeak + slp + caa + thall, family = binomial, data = trainingdata)
##
## Coefficients:
## (Intercept)         sex1         cp.L         cp.Q         cp.C       trtbps
##     3.55136     -1.79970      2.15369     -0.38044     -0.06896     -0.02343
##    thalachh      oldpeak        slp.L        slp.Q          caa      thall.L
##     0.01705     -0.43681      0.28953      0.92982     -0.70652      0.02574
##     thall.Q      thall.C
##    -1.94784      0.96015
##
## Degrees of Freedom: 239 Total (i.e. Null);  226 Residual
## Null Deviance:       331.9
## Residual Deviance: 170.1     AIC: 198.1
```

Bidirectional elimination resulted in the same model as the backward elimination model.

4. The Best Logistic Model

The two logistic models will be assessed against the remaining 25% data, the training data, to see which model has more predictive power.

A. Backward Elimination/ Bidirectional Elimination

```
##       Predicted
## Actual  0  1
##      0 18  7
##      1  5 33
```

```
## Accuracy = 0.81
```

B. Orginal/ Forward Selection

```
##       Predicted
## Actual  0  1
##      0 19  6
##      1  5 33
```

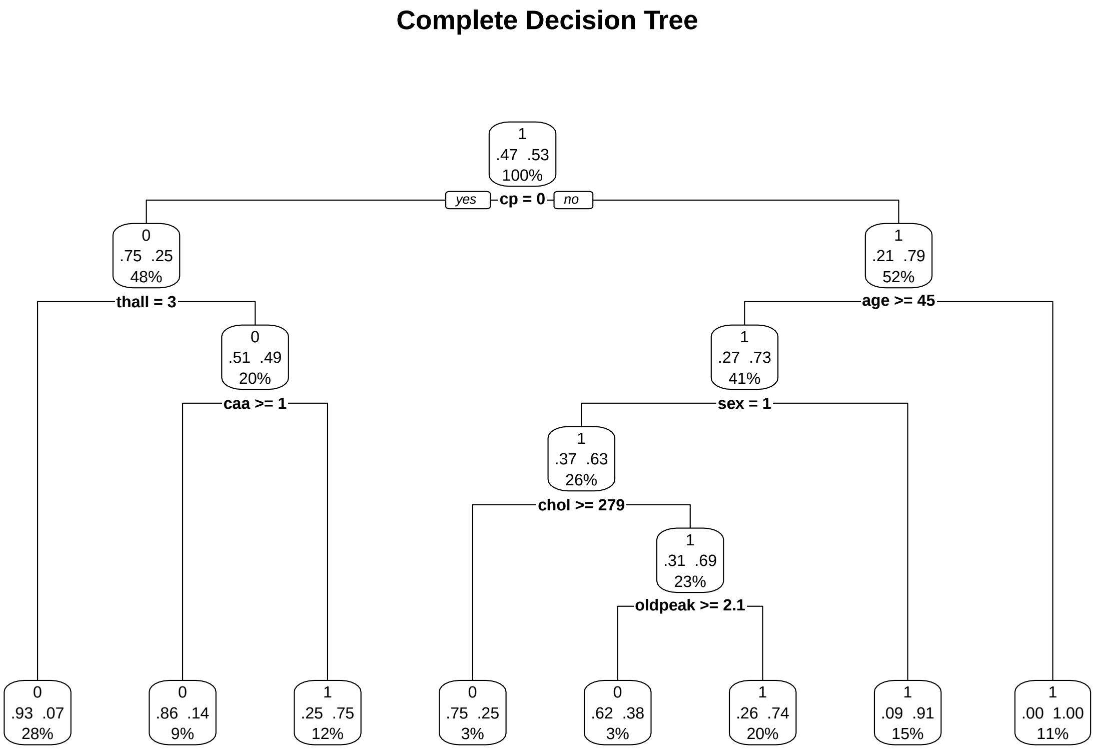
```
## Accuracy = 0.83
```

This is a tough choice as both regression models excel at different areas. While the backward elimination regression model has a slightly lower AIC, it had slightly lower predictive power. The backward elimination model selected the wrong classification 11 times or 17% percent of the time compared to the original model which elected the wrong classification 10 times. As the differences in AIC and predictive power where almost negliblle, the original model was kept.

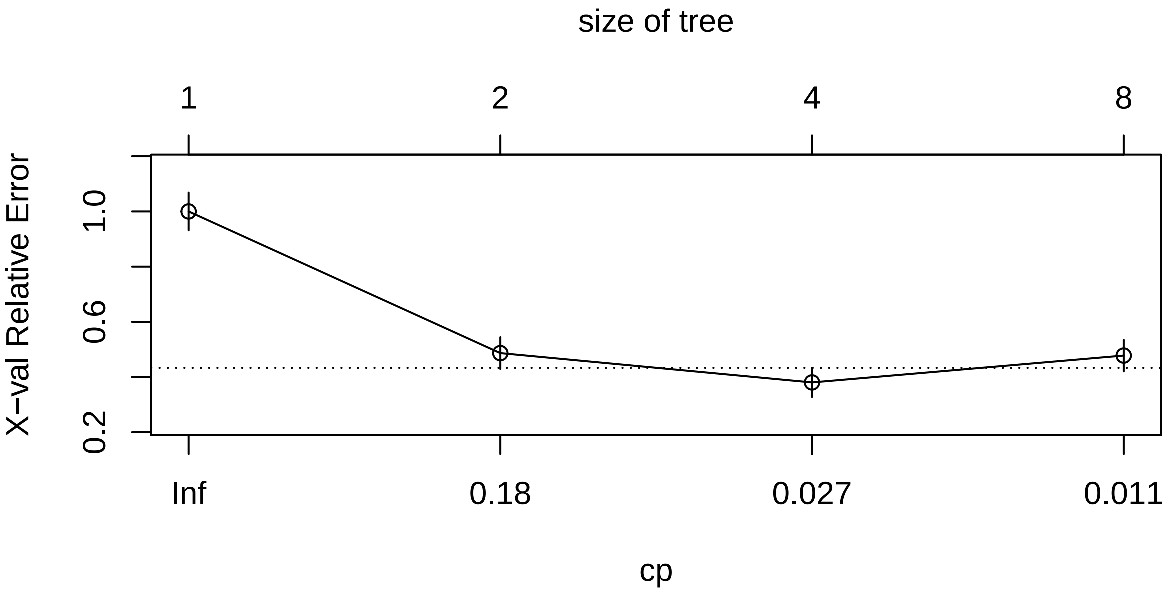
Classical Decision Trees

1. Complete Tree

Below is a graphical representation of the Complete decision tree and CP (Complexity parameter) table and plot.

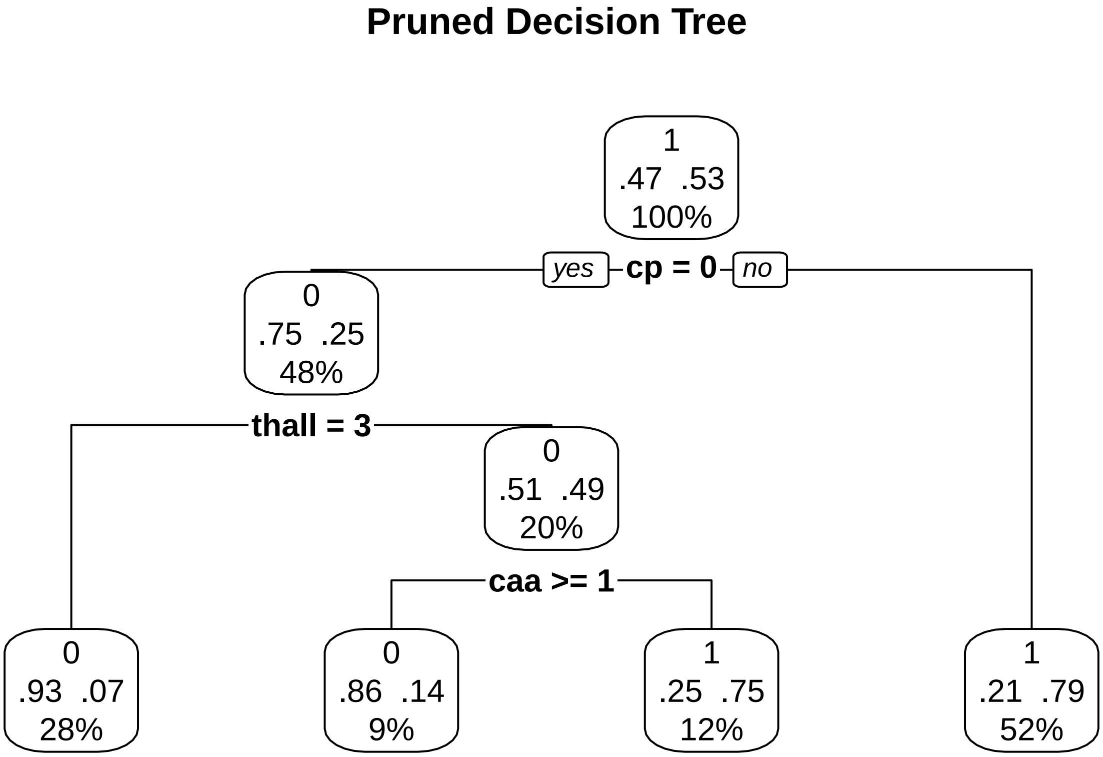


```
##           CP nsplit rel error    xerror       xstd
## 1 0.51327434      0 1.0000000 1.0000000 0.06843165
## 2 0.06194690      1 0.4867257 0.4867257 0.05762131
## 3 0.01179941      3 0.3628319 0.3805310 0.05257548
## 4 0.01000000      7 0.3097345 0.4778761 0.05724918
```



2. **Pruned Tree**

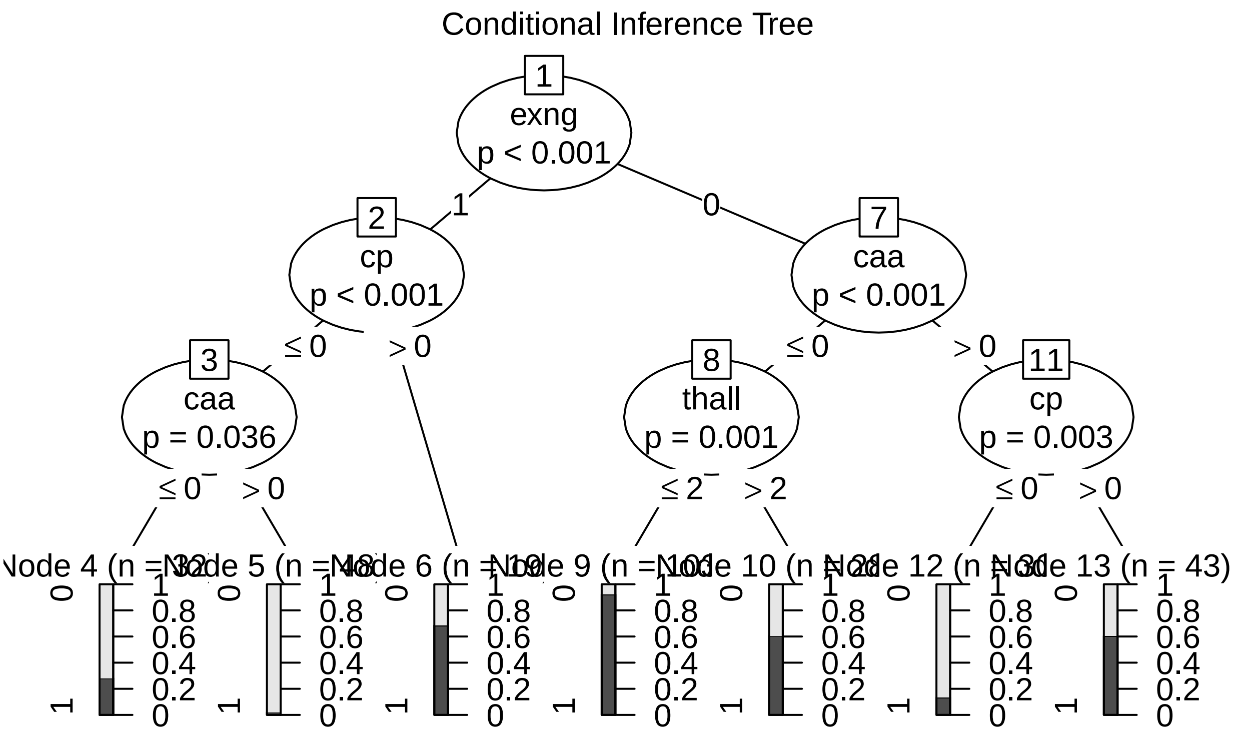
Based from the CP table and plot, the decision tree will be pruned at the 3 nsplit or where the CP = .027. Below is the graph of the pruned decision tree and its classification of the testing data.



```
## Accuracy = 0.81
```

Conditional Inference Trees

Below is the graphical representation of Conditional Inference Tree and its classification of the testing data.



```
##       Predicted
## Actual  0  1
##      0 16  9
##      1  1 37
```

Random Forest

Below is the code output of Random Forest and its classification of the testing data.

```
##
## Call:
##  randomForest(formula = formulaf, data = trainingdata, importance = TRUE,      na.action = na.roughf
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 20%
## Confusion matrix:
##    0   1 class.error
## 0 84  29   0.2566372
## 1 19 108   0.1496063
```

```
##          MeanDecreaseGini
## age             10.143370
## sex              4.276555
## cp              19.335035
## trtbps           9.418173
## chol             9.910214
## fbs              1.138912
## restecg          2.517282
## thalachh        12.735692
## exng             6.058946
## oldpeak         12.156288
## slp              6.803451
## caa             11.442321
## thall           11.998021
```

```
##       Predicted
## Actual  0  1
##      0 20  5
##      1  2 36
```

Assesing Classification Accuracy

1. Performance of the Logistic Regression

```
## Sensitivity = 0.87
## Specificity = 0.76
## Positive Predictive Value = 0.85
## Negative Predictive Value = 0.79
## Accuracy = 0.83
```

2. Performance of the Classical Decision Tree

```
## Sensitivity = 0.95
## Specificity = 0.6
## Positive Predictive Value = 0.78
## Negative Predictive Value = 0.88
## Accuracy = 0.81
```

3. Performance of the Conditional Inference Tree

```
## Sensitivity = 0.97
## Specificity = 0.64
## Positive Predictive Value = 0.8
## Negative Predictive Value = 0.94
## Accuracy = 0.84
```

4. Performance of the Random Forest

```
## Sensitivity = 0.95
## Specificity = 0.8
## Positive Predictive Value = 0.88
## Negative Predictive Value = 0.91
## Accuracy = 0.89
```

Conclusion

While all models had a relatively high predictive power, with each model having at least an 80% accuracy. The Random Forest had the most accuracy, with it being accurate almost 90% of the time. However, it is important to note that while it is the most accurate it is the least interpretable of the models. If a more interpretable model was necessary a conditional inference tree could be selected instead. And this could make sense in a medical setting as patients might be worried about how the model diagnosed them.

References

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.