# Predicting Heart Disease Dataset by SVM and Decision Trees

By Paulette Rodriguez

# Introduction

- Commonly known as Cardiovascular Disease (CVD), heart diseases are groups of disorders in the heart and blood vessels.
- CVDs are the number one cause of death globally, taking an estimated 17.9 million lives each year, according to the World Health Organization (WHO).
- Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States with one person dying every 37 seconds in the U.S.
- Statistically speaking that's about 647,000 Americans dying from heart diseases each year. That's a ratio of 1 in every 4 deaths.

# The Dataset

The Heart Disease dataset being used for this analysis project consists of 303 individual's data. It contains 76 attributes, using a subset of only 14 of them.

It can be found here:
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Credited to:

| | | | |
|---|---|---|---|
| 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. |

# The Attributes Evaluated

Age

Sex

Chest-Pain (cp)

Resting Blood Pressure (rbp)

Serum Cholestorol (chol)

Fasting Blood Sugar (fbs)

Resting ECG (restecg)

Thalach

Exang

Oldpeak

Slope

Fluoroscopy (ca)

Thal

Num

# Approach

THE USE OF R PROGRAMMING WILL BE USED TO IMPLEMENT DIFFERENT CLASSIFICATION MODELS ON THE DATASET.

THE FOLLOWING CLASSIFICATION MODELS THAT WILL BE IMPLEMENTED ON THE SET WILL BE:

SVM

DECISION TREES

These different classification models will be used to help predict accuracy and useful output in determining the odds of getting heart disease based on different risk factors.

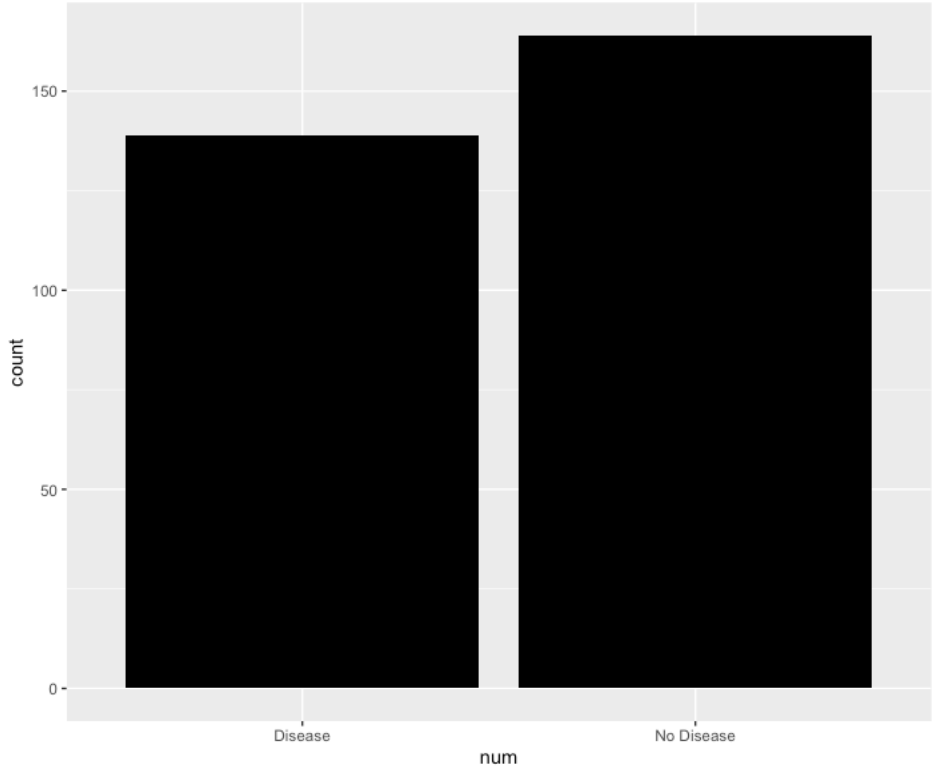
# Analysis & Results

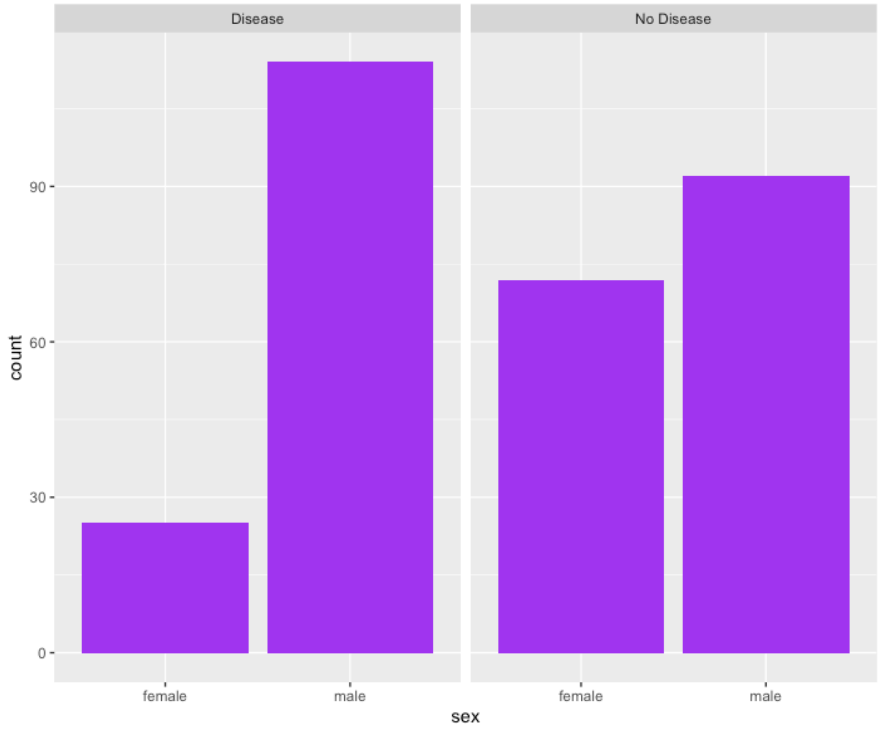
- The following output shows that 139 individuals have heart disease, while 164 do not.

| Disease | No Disease |
| --- | --- |
| 139 | 164 |

# Analysis/Results (CONT'd)

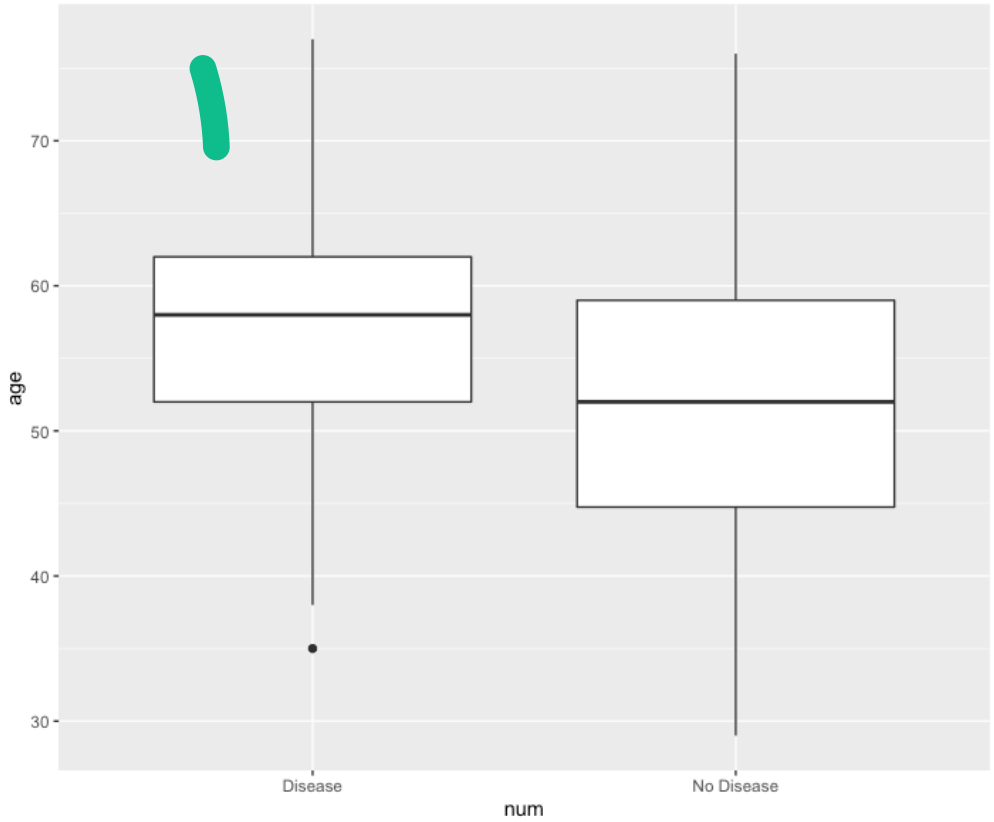
- By showing the distribution of the "num" variable, against the gender variable, we can see the numbers plotted nicely for us to understand the data better.



# Analysis/Results (CONT'd)

- Also, by making a box plot of "num" and "age", we can understand the statistical distribution between the two variables. As a result, we can see that the individuals who did have heart disease were around an average age of 56.



# Analysis/Results (CONT'd)

- Now, let's run some correlation analysis between "age" and "chol".

```
cor.test(age, chol)
```

```
##
##  Pearson's product-moment correlation
##
## data:  age and chol
## t = 3.707, df = 301, p-value = 0.0002496
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09859353 0.31423005
## sample estimates:
##       cor
## 0.2089503
```

- We can see that age and cholesterol levels have very low correlation

# Analysis/Results (CONT'd)

- Now let us output a confusion matrix of chest pain and heart disease as well as a confusion matrix of exercise induced asthma and heart disease.

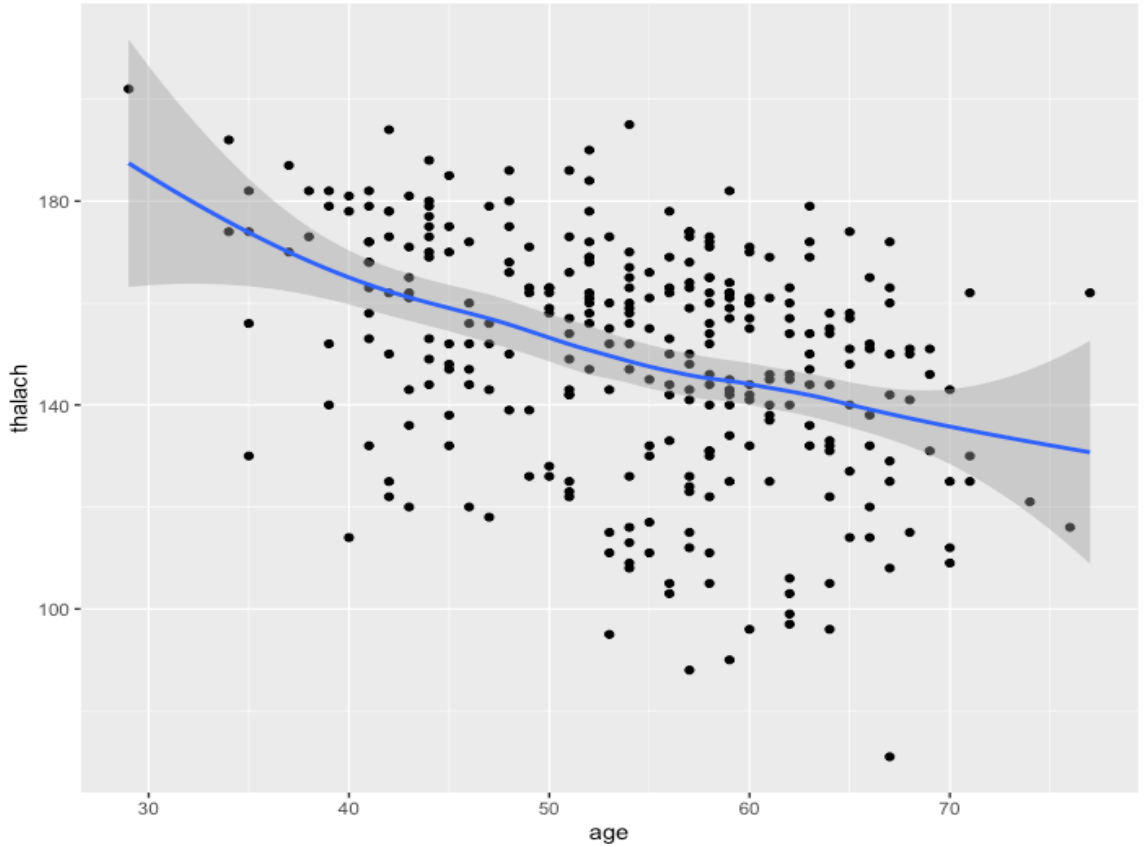
| | Num | | | | |
|---|---|---|---|---|---|
| Cp | 0 | 1 | 2 | 3 | 4 |
| 1 | 16 | 5 | 1 | 0 | 1 |
| 2 | 41 | 6 | 1 | 2 | 0 |
| 3 | 68 | 9 | 4 | 4 | 1 |
| 4 | 39 | 35 | 30 | 29 | 11 |

| | Num | | | | |
|---|---|---|---|---|---|
| exang | 0 | 1 | 2 | 3 | 4 |
| 0 | 141 | 30 | 14 | 12 | 7 |
| 1 | 23 | 25 | 22 | 23 | 6 |

- By doing so, we will be able to notice that the individuals who had heart diseases had severe levels of chest pain and exercise induced asthma.

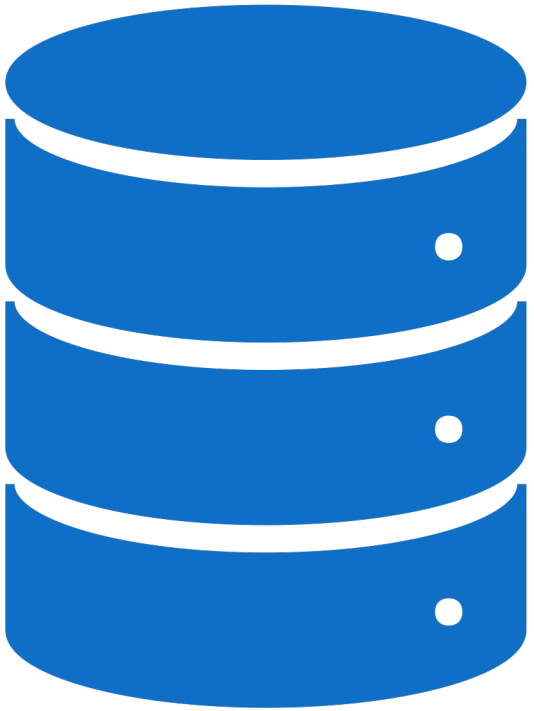
# Analysis/Results (CONT'd)

- As age increases, maximum heart rate achieved decreases since the correlation between the two is negative.



# SVM

- SVM classifier tends to generate maximum marginal hyperplanes.
- By using the SVM classifier, our main goal will be to separate the classes on either side of the hyperplane with maximum margin using the support vectors.
  - For example, a linear SVM classifier will generate a simple linear hyperplane for linearly separable data.



# SVM

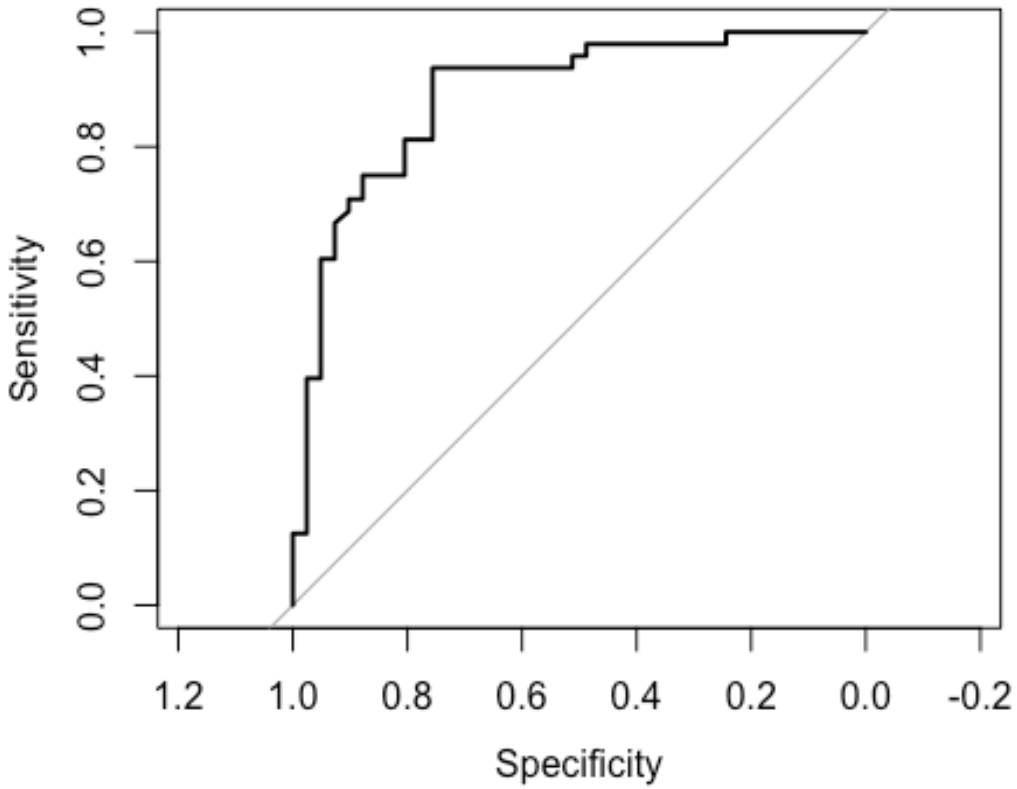
- As a result of using SVM for prediction of our data, we got the following outputs:

**AUC and ACURACY**

| | AUC | Accuracy |
|---|---|---|
| Accuracy | 0.890498 | 0.8314607 |

Hence, we get an AUC value of about 0.89% and an overall prediction accuracy value of 0.83%.

**ROC CURVE**

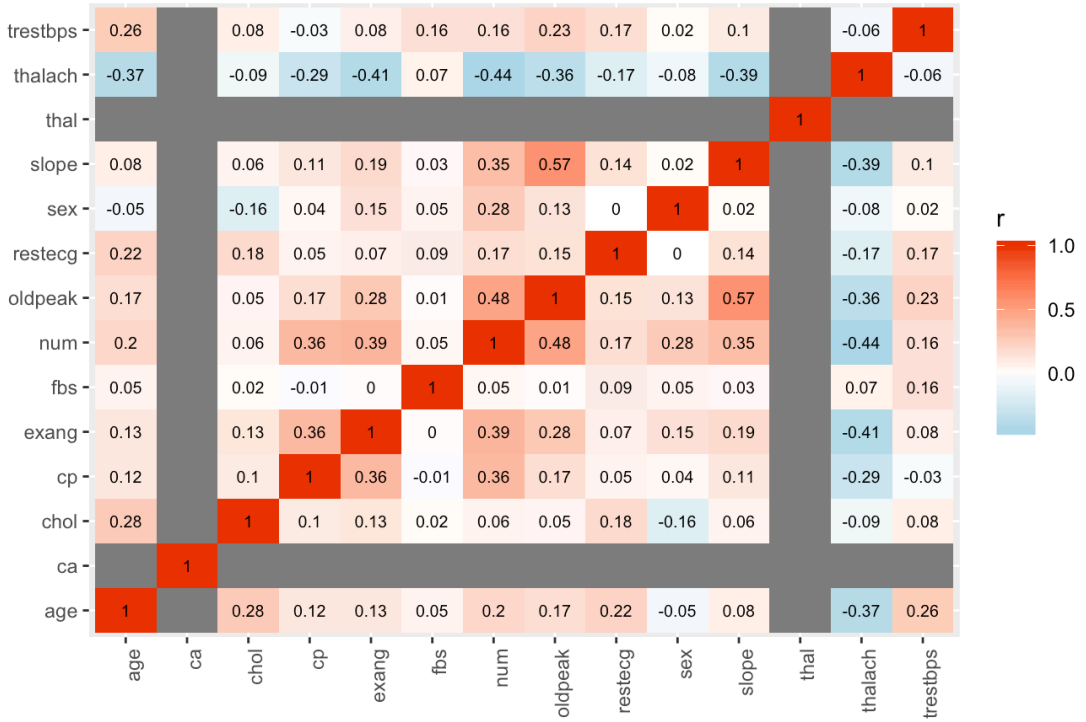


# Decision Trees

- Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems.
- There are different types of decision trees such a:
  - Classification Trees
  - Regression Trees
- A classification tree is very similar to a regression tree, except that it is used to predict a qualitative response rather than a quantitative one.

# Decision Trees (CONT'd)

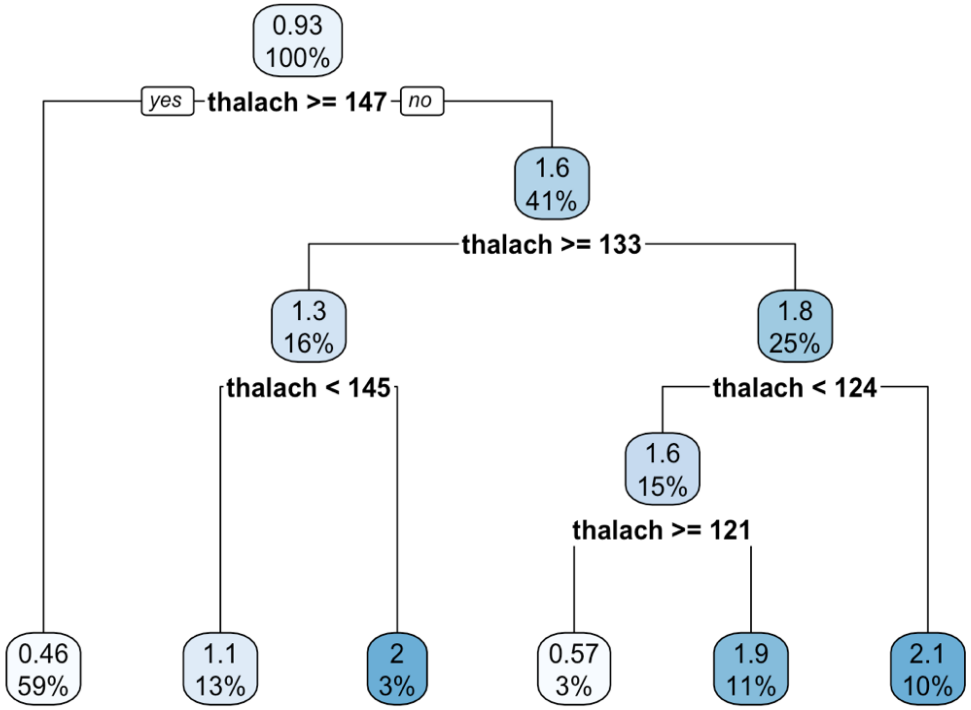
- Looking at the correlation of the data, we get the following correlation matrix:

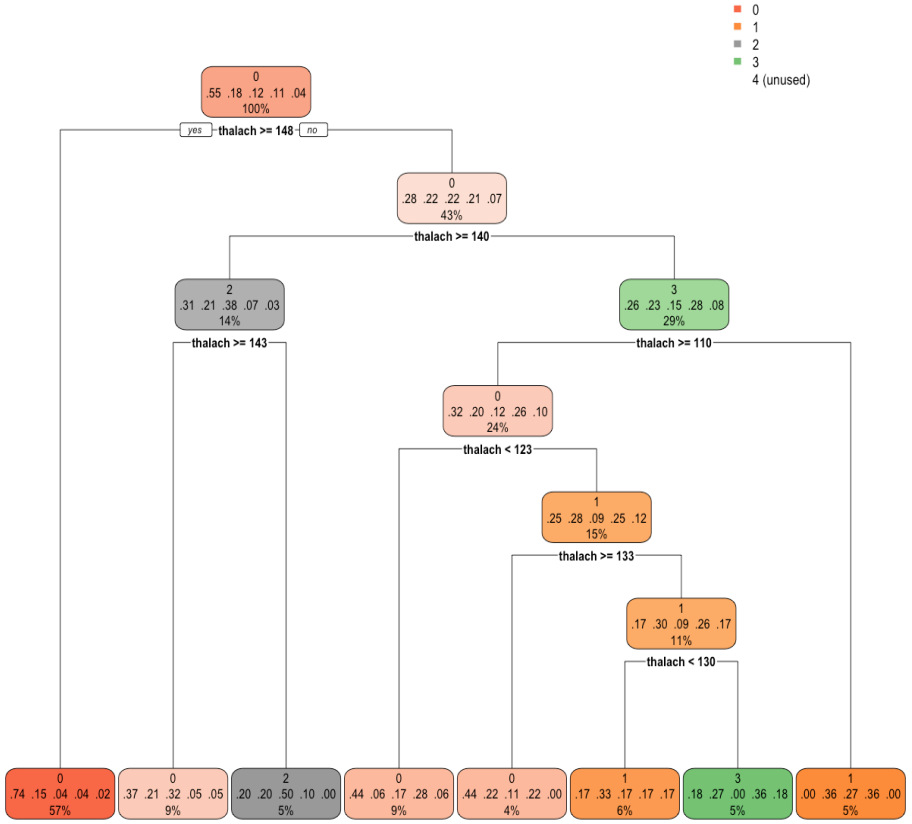



- Our variable "thalach" has the greatest negative correlation with "num."

# Decision Trees (CONT'd)

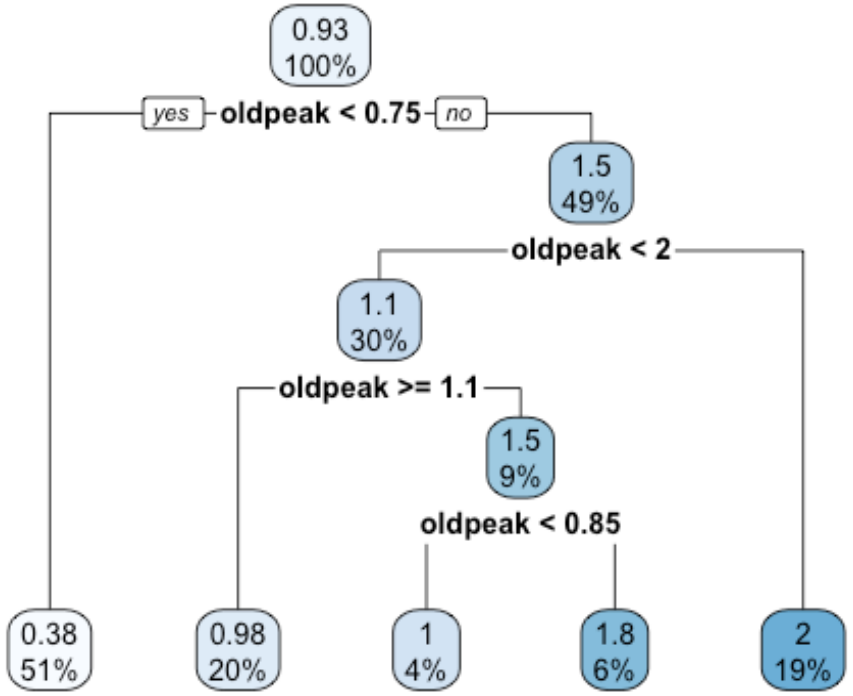
- Given the information from the previous slide, we will start to build a regression and classification tree (respectively) using thalach.

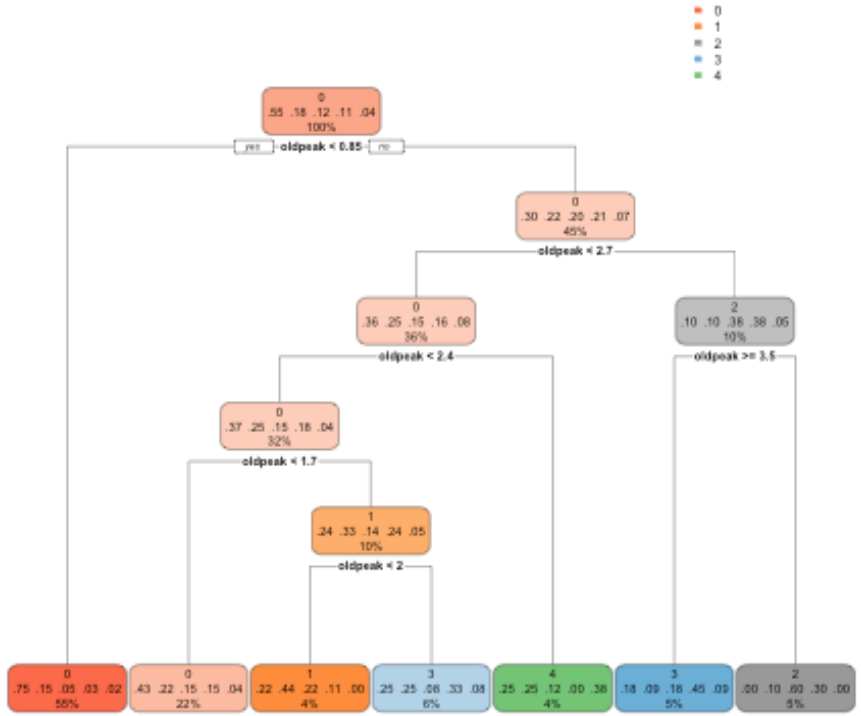


# Decision Tree (CONT'd)

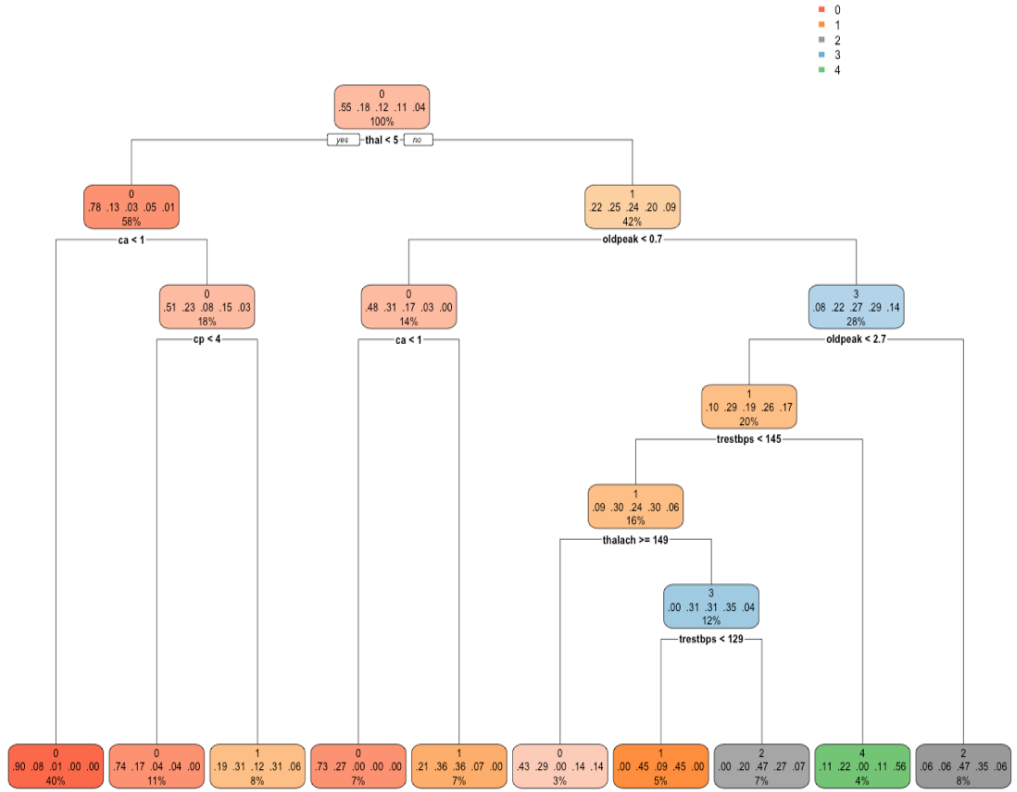
- As a result of the correlation matrix we built, we were also able to see that our variable "oldpeak" had the greatest postive correlation with "num." Let's look at a regression and classification tree (respectively) using "oldpeak."

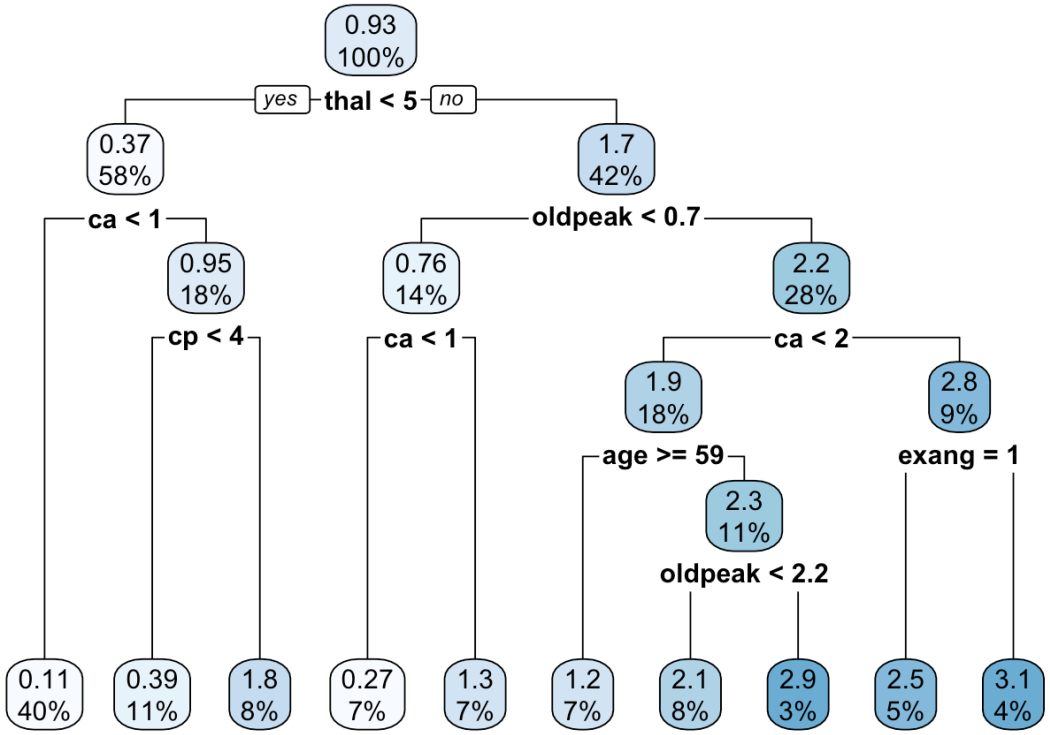


# Decision Trees (CONT'd)

- Now that we have looked at regression and classification trees using our variables that gave us the greatest positive and negative correlation response, we can look at a classification tree and regression tree using all the variables in our data.



# Model Comparison

- We see that the highest accuracy for the test set is achieved by SVM, which gave us a result of 89% for AUC and 84% for prediction accuracy.
- The highest accuracy for the training set is archived by Decision Tree which, as we can see from our output, turned out to be a result of 100% which is high.

# Limitations

- Decision trees for regression and classification have several advantages and disadvantages.

**Advantages:**

- Trees are very easy to explain. They are in fact easier to explain than linear regression.
- Some people believe that decision trees more closely mirror human decision-making techniques than some of the regression and classification approaches we have seen in our lives.
- Trees can be displayed graphically, which makes it a lot easier to interpret even by people who are not experts.
- Trees can easily handle qualitative predictors without the need to create dummy variables.

**Disadvantages:**

- trees generally do not have the same levels of predictive accuracy as some other regression and classification approaches that we might have worked with before in our lives
- However, with methods such as bagging, random forests, and boosting, the predictive performance of trees can be substantially improved.

# Limitations (CONT'd)

- Just as the decision tree algorithm has its disadvantages, so does SVM.

**Advantages:**

- SVM works relatively well when there is a clear separation in the margin between classes.
- It is more effective in high dimensional places.
- It is effective in cases where the number of dimensions is greater than number of the samples.
- It is memory efficient.

**Disadvantages:**

- SVM algorithm is not suitable for large data sets.
- SVM does not perform well when the data has target classes that overlap.
- SVM will underperform in cases where the number of features for each data point is larger than the number of training data samples.

# Conclusion

- Throughout this analysis and model comparison, we used 14 predictor variables from the heart disease dataset to predict whether a patient had presence of heart diseases.
- We used different models such as SVM which gave us a high accuracy of 89% and 83%.
- Then, we used Decision Trees to visualize the data in a better way. With Decision Trees, we were able to see an output of a high accuracy from that as well.
- Hence, we can conclude that from the heart disease data results the best variables that show presence of heart disease are “num”, “cp”, “ca”, and “thal.”