Paulette Rodriguez
MATH 678
Project Report
9 May 2020

Predicting Heart Disease Dataset by SVM and Decision Trees

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). The different groups of heart disease vary from coronary heart disease, cerebrovascular disease, rheumatic heart disease and others. According to the Centers of Disease Control and Prevention (CDC), 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.

Not only do CVDs take millions of lives every year, but it turns out that it's super expensive and costs the United States over $200 billion each year. For example, the CDC has stated that heart disease cost the United States about $219 billion each year from 2014 to 2015. Which included the cost of health care services, medicines, and lost productivity due to death.

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. The "goal" field refers to the presence of heart disease in the patient. The data is integer valued from 0, meaning no presence of heart disease, to 4 with number 1-4 distinguishing presence of the disease. As stated above, the dataset contains only 14 of the 76 total attributes. These following 14 attributes are described below:

1. **Age:** displays the age of the individual in years.
2. **Sex:** displays the gender of the individual using the following format:
   1 = Male
   2 = Female
3. **Chest-pain type (cp):** Displays the type of chest-pain experienced by the individual.
   1 = typical angina
   2 = atypical angina
   3 = non-anginal pain
   4 = asymptotic
4. **Resting Blood Pressure (RBP):** displays the resting blood pressure value of an individual in unit mmHg.

5. **Serum Cholesterol (chol):** displays the serum cholesterol in unit mg/dl
6. **Fasting Blood Sugar (fbs):** compares the fasting blood sugar value of an individual with 120 mg/dl.
   If: fasting blood sugar > 120 mg/dl, then: 1 (true), else: 0 (false).
7. **Resting ECG (restecg):** displays resting electrocardiographic results.
   0 = normal
   1 = having ST-T wave abnormality
   2 = left ventricular hypertrophy
8. **Max hear rate achieved (thalach):** displays the max heart rate achieved by an individual.
9. **Exercise induced angina (exang):**
   1 = yes
   2 = no
10. **ST depression (oldpeak):** displays the value which is an integer or float of ST depression induced by exercise relative to rest.
11. **Peak exercise ST segment (slope):**
   1 = upsloping
   2 = flat
   3 = down sloping
12. **Fluoroscopy (ca):** displays the value as integer or float of number of major vessels (0-3) colored by fluoroscopy.
13. **Thal:** displays the maximum heart rate achieved.
14. **Diagnosis of heart disease (num):** displays whether the individual is suffering from heart disease or not:
   0 = absence
   1,2,3,4 = present.

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 and 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.

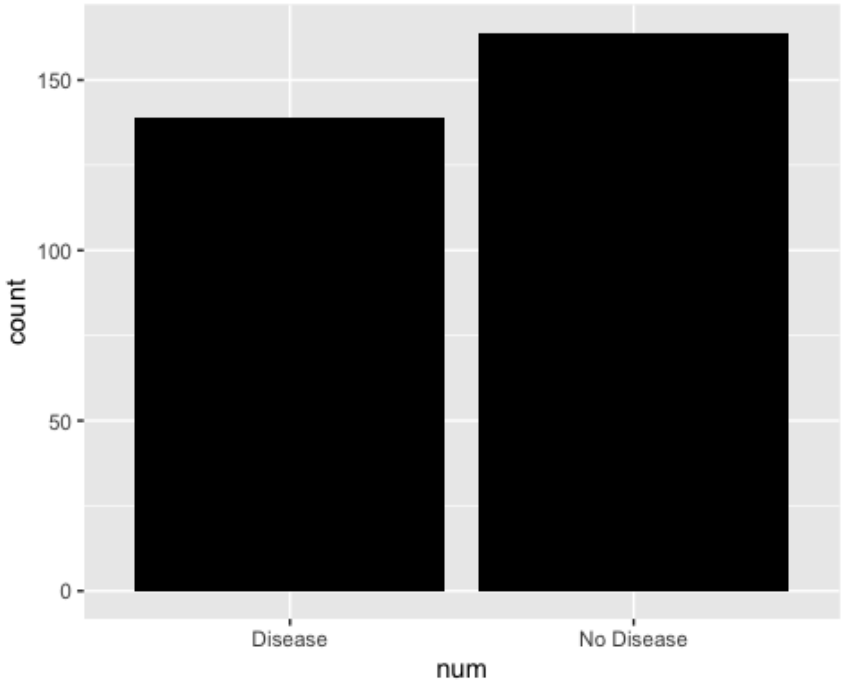
Analysis & Results:

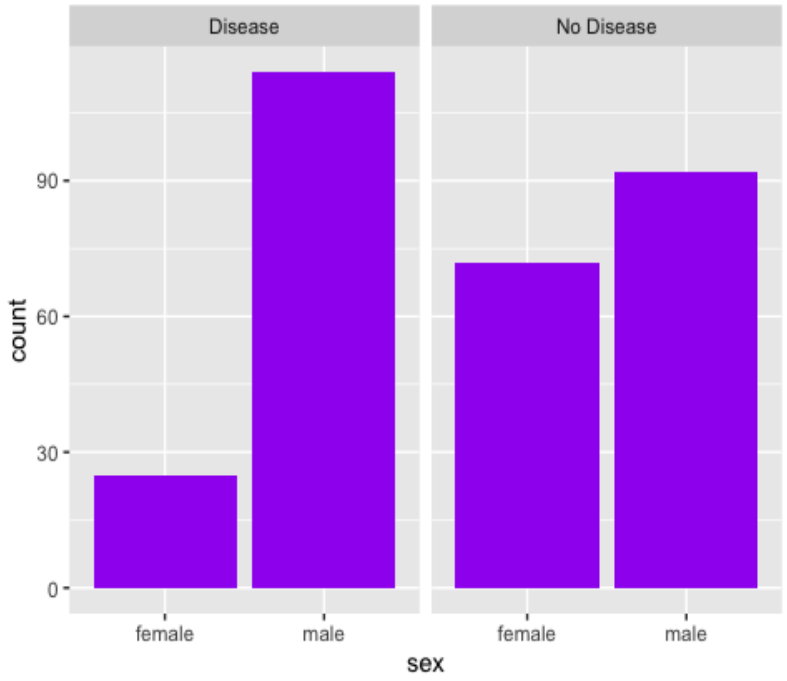
In the following steps, I will be using SVM and Decision Trees classification techniques to help predict heart disease. From the set of 14 attributes we saw above, there are more important variables that come into play in the prediction of an individual getting heart disease. The most important variables to predict heart failure in an individual come from whether or not there is a reversable defect in Thalassemia followed by whether or not there is an occurrence of chest pain. Please note, target = 1,2,3,4 implies that the person is suffering from heart disease and target = 0 implies that the given individual is not suffering from the disease. The variable we want to predict is "num" with a value 0 being less than 50% and value 1 being greater than 50%.

Given that the distances between the values are random, such as “cp”, “thal”, “restecg”, and “slope”, it is clear that we need to dummify these variables. Now, let’s take a quick view of the data and explore it to find how many individuals had hearts attacks based on gender and age. First, we will convert the dependent variable “num” to a binary variable. The following output shows that 139 individuals have heart disease, while 164 do not.

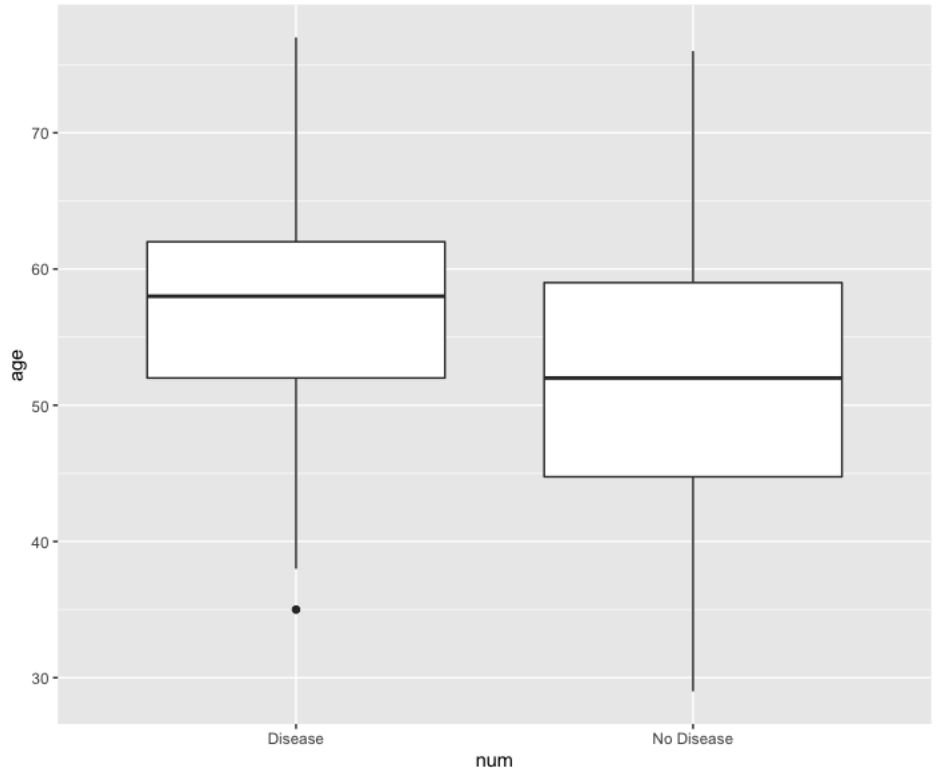
Disease	No Disease
139	164

By showing the distribution of the “num” variable, against the gender variable, we are able to see the numbers plotted nicely for us to understand the data better.





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



age of 56.

Now, let's do some correlation analysis between some of the variables. For example, 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 are able to see that age and cholesterol levels have very low correlation.

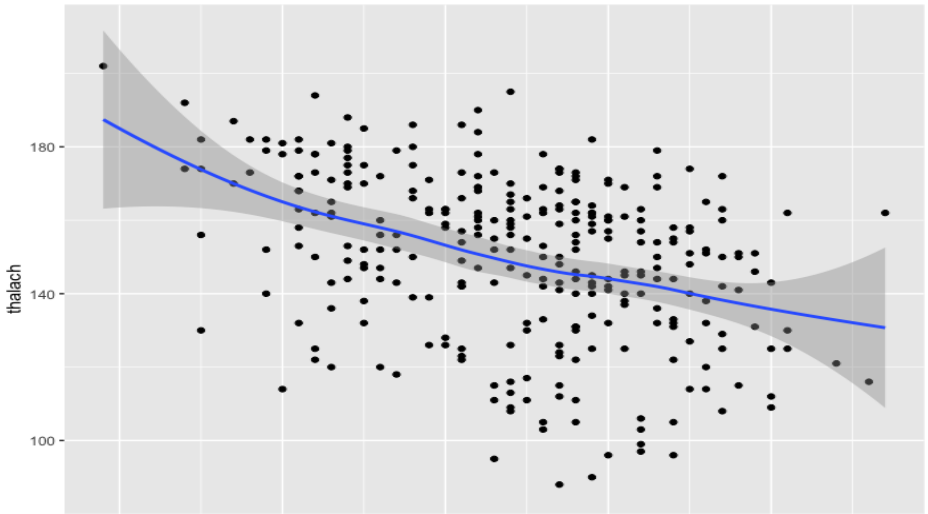
Now let us output a confusion matrix of chest pain and heart disease as well as a confusion matrix of exercise induced asthma and heart disease.

	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. Also, that those individuals who had heart disease had asthma induced from exercising.

Hence, the correlation between age and maximum heart rate is achieved. We can see from the graph shown that as age increases, maximum heart rate achieved decreases since the correlation between the two is negative.



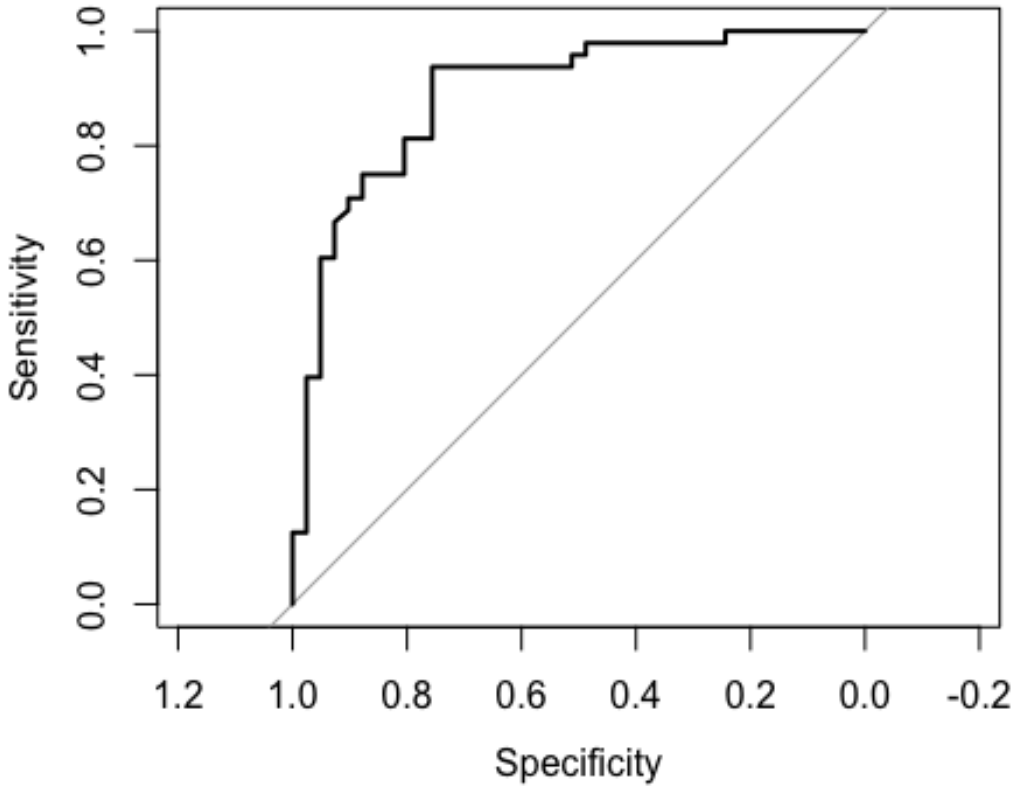
SVM:

Before getting into building an SVM classifier we will split the data into training and testing datasets. Now, the 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. For this to work we can't allow any numbers when adding the names of the data to all levels. This will lead us into feature selections as shown in the given code. This will help us when converting to factor variable with only two levels, "Disease" or "No Disease." After forming our SVM model we are able to then predict on the test data class labels, find the probability of no heart disease, and generate a confusion matrix. The confusion matrix, as shown below, has a diagonal that represent the correctly classified labels, and the off diagonals which are incorrect classifiers.

	Reference	
Prediction	Not Disease	Disease
Not Disease	31	5
Disease	10	43

Now, to better understand the accuracy and the performance of the SVM classifier we will find the ROC curve and the area under the curve (AUC) value. The ROC curve is the plot of true positive rates versus the false positive ones. By running the code, we get the following output.

	AUC	Accuracy
Accuracy	0.890498	0.8314607

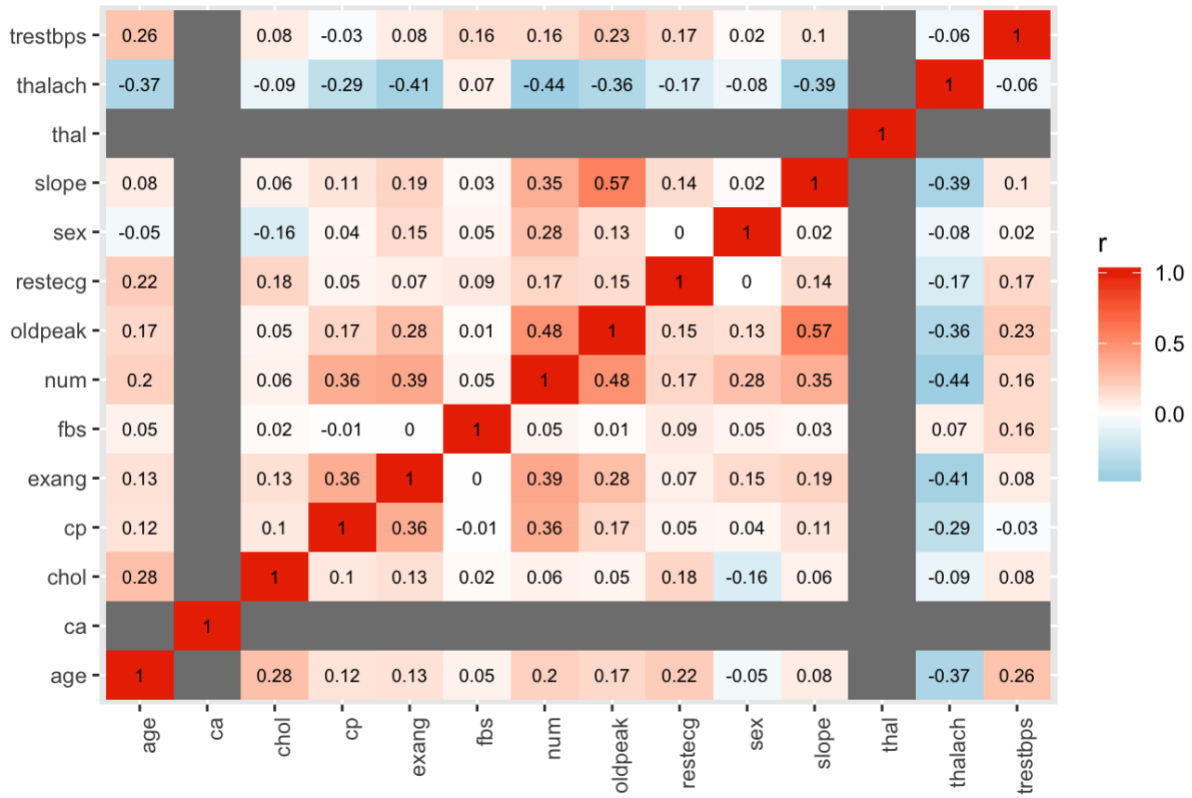


Hence, we get an AUC value of about 0.89 and an overall prediction accuracy value of 0.83.

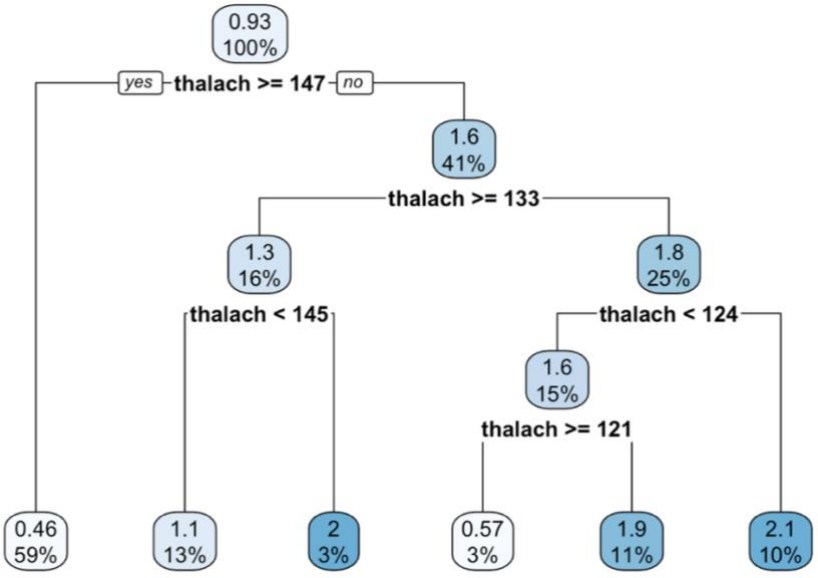
Decision Trees:

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 as classification trees and 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. For this analysis, we will take a look at both a regression tree and a classification tree.

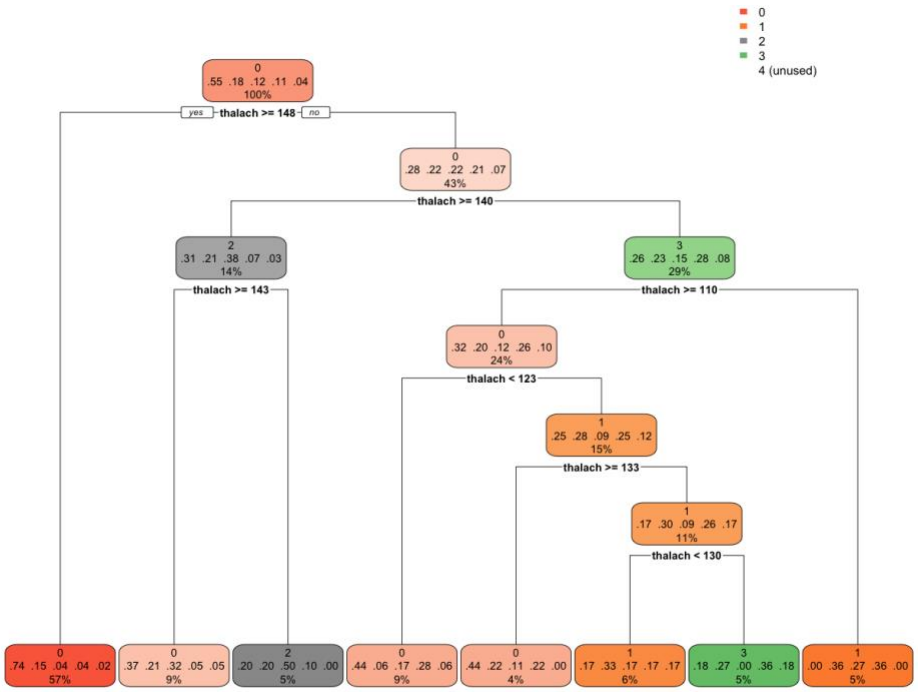
Just as we did before, we will start by importing the data as the first step followed by exploratory analysis. Then, we need to split the sample data for training and testing purposes. After we do that, we will take a look at the correlation of the data. This is similar to how we found the correlation for our SVM analysis. As a result, we get the following correlation matrix.



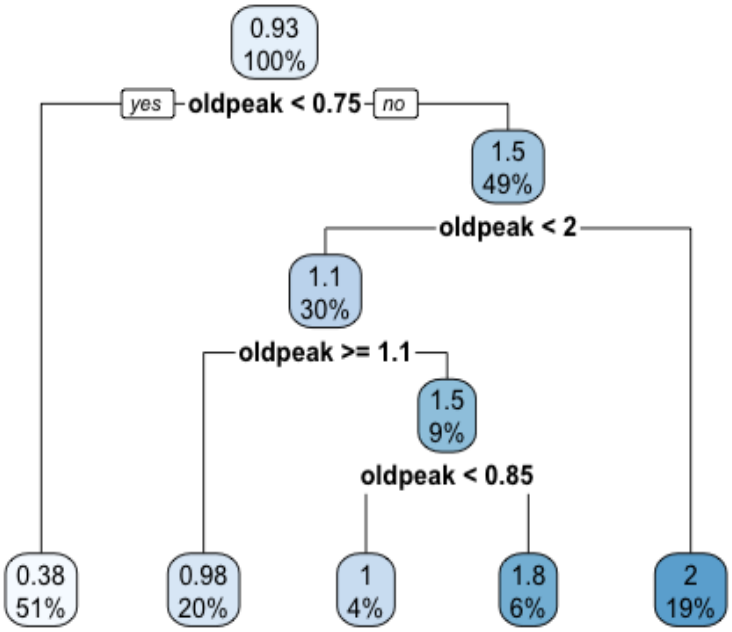
Our variable "thlach" has the greatest negative correlation with "num", which as we saw before is the variable that shows the presence of heart disease in the individual. Given this information, we will start to build a regression tree using thalach.



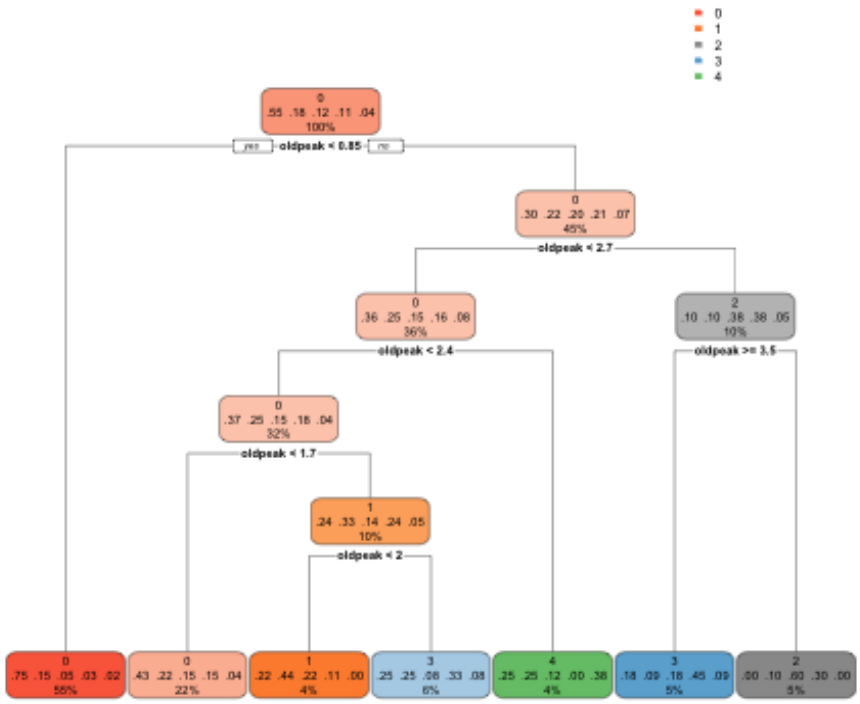
Then, we can also build a classification tree using thalach.



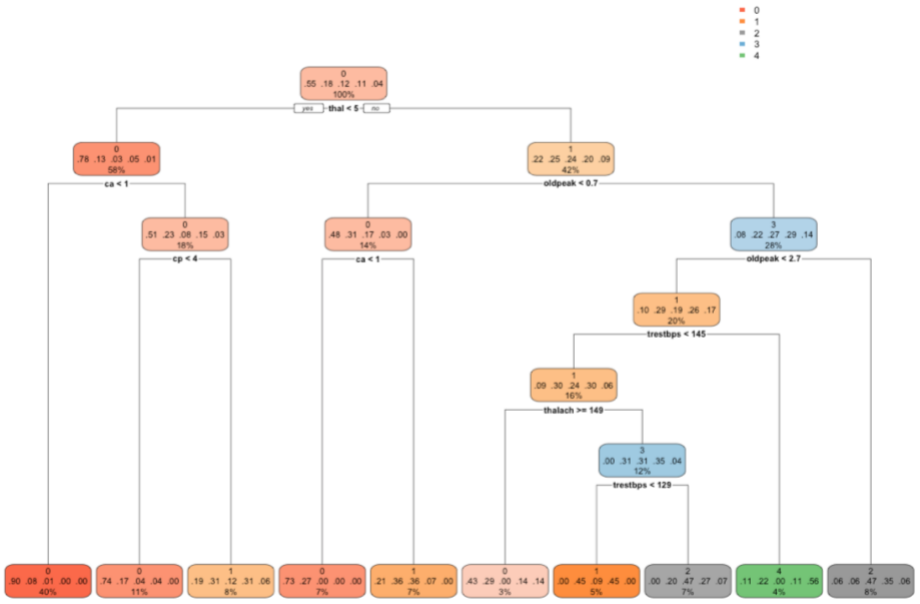
As a result of the correlation matrix we built, we were also able to see that our variable "oldpeak" had the greatest psotive correlation with "num." Let's take a look at a regression tree using "oldpeak."

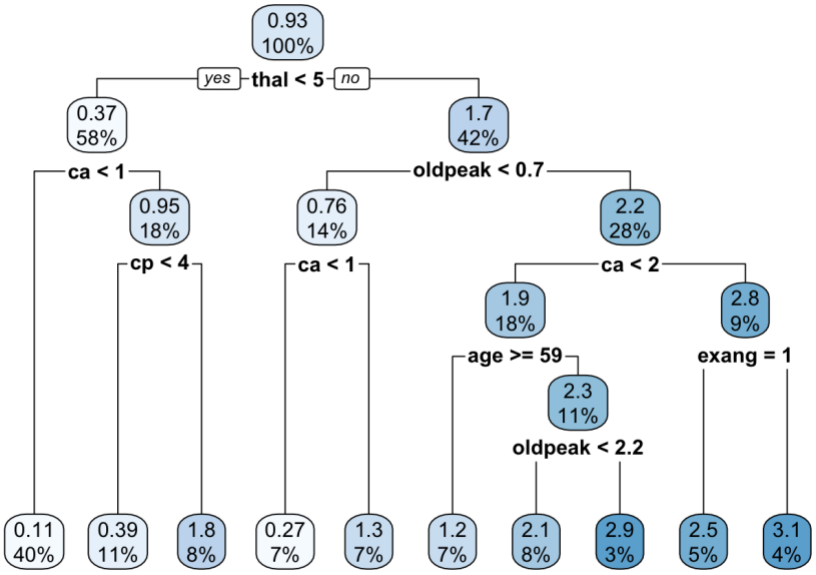


Now, a classification tree.



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





Model Comparison:

Now, to sum up our predictions with using SVM and Decision trees we will take a look at the accuracies that we achieved with the models.

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 really high.

By using both of these algorithms, we were able to predict pretty good accuracies for both the training and the testing sets that we defined.

Limitations:

Decision trees for regression and classification have a number of advantages and disadvantages. For example, some of the advantages include the following:

- 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.

Unfortunately, trees do have some disadvantages. For example, one disadvantage of decision trees is that 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.

Just as the decision tree algorithm has its disadvantages, so does SVM. For example, some of the advantages that SVM has include the following:

- 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.

Unfortunately, SVM does have some disadvantages of its own. For example, some of its disadvantages include the following:

- 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 or not a patient had presence of heart diseases. In order to do this, we used different models. The first model we saw was SVM which gave us a pretty high accuracy of 89% and 83%. Then, we used Decision Trees to visualize the data in a better way. Using the method of 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.”

Heart disease is one of the major concerns for people all around the world till this day. With the help of some useful techniques that were used for this analysis, such as SVM and Decision Trees, we were able to predict the odds of getting heart disease based on various risk factors. It helped a lot that these techniques were done using algorithms on a computer, since doing some analysis like this would have be difficult to achieve manually.

References:

1. https://www.academia.edu/36691506/An_Introduction_to_Statistical_Learning_Springer_Texts_in_Statistics_An_Introduction_to_Statistical_Learning
2. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
    a. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
    b. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
    c. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
    d. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.
3. https://www.cdc.gov/heartdisease/facts.htm
4. https://www.datacamp.com/community/tutorials/decision-trees-R
5. https://dzone.com/articles/support-vector-machine-in-r-using-svm-to-predict-h
6. https://github.com/Elmar999/Heart_disease_project/blob/master/Report.pdf
7. https://www.kaggle.com/wguesdon/predicting-heart-disease-risk-with-random-forest
8. https://rpubs.com/mbbrigitte/heartdisease
9. https://rpubs.com/ssshinde/Heart_disease_prediction
10. https://towardsdatascience.com/heart-disease-prediction-73468d630cfc
11. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1