A Comparative Assessment of Machine Learning Models on Heart Disease Data

December 1, 2023

Alexander Hernandez, Salik Faisal, Angel Rodriguez, Ariana Quilantan

Introduction

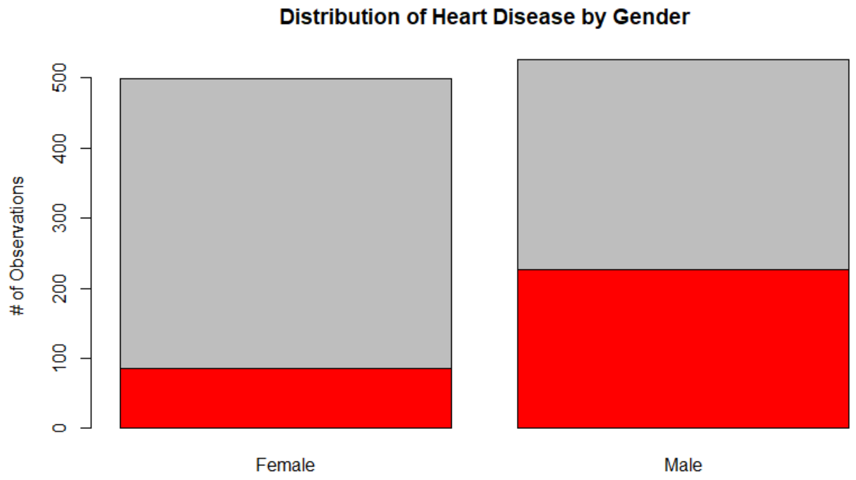
Heart disease is the leading cause of death for men and women in the United States. According to the CDC, one person dies of heart disease, or cardiovascular disease every 33 seconds (1), and costs the United States nearly $240 billion due to the cost of healthcare, medicine, and loss of productivity (2). It is safe to say then, that we each know or have known someone with heart disease at some point in our lives, and despite our relative perception of health, it is something that can affect any of us. We have therefore decided to conduct this report based on data regarding heart disease. Using the data described below, we wish to understand which variables are the best predictors when it comes to accurately predicting our response variable.

Dataset Information (Angel Rodriguez)

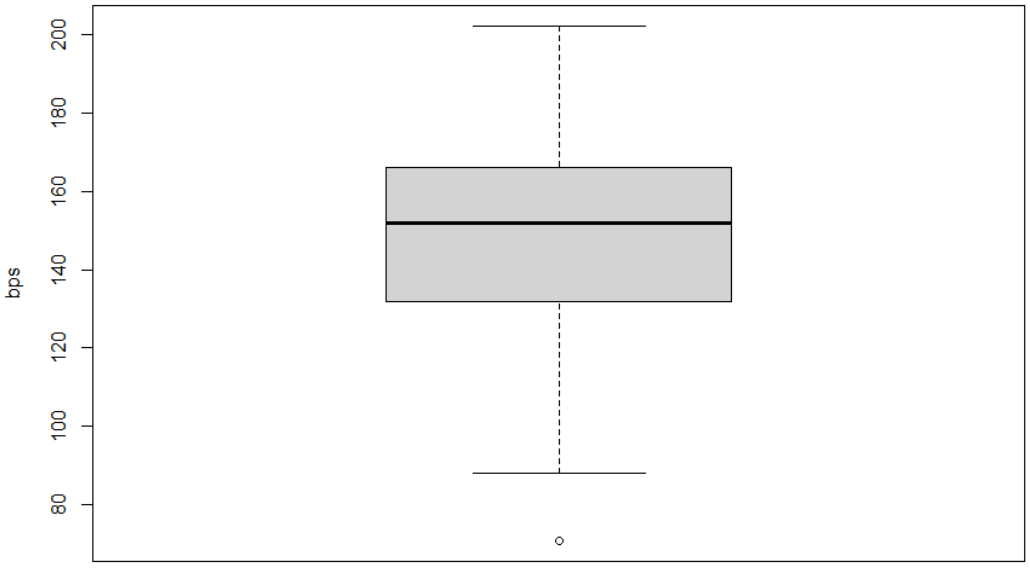
The data was obtained from a website called Kaggle, a data science community page that allows users to find data sets they want to use for the purpose of training or creating data science and machine learning models. The data set used is called "Heart Disease Dataset". The information here is recorded from 1025 patients in 1988. Each patient had 14 variables recorded. The first variable is **Age**, a continuous variable showing the age of the patient. The second variable is the binary variable **Sex**. This variable shows whether a patient is a Female (1) or Male (0). The third variable is **CP** (Chest Pain) and it is a discrete variable ranging from 0-4 showing what type of chest pain the patient was experiencing. The fourth variable is **Chol**, which measures the serum cholesterol in the patient's blood in mg/dl. This variable is a continuous variable. The fifth variable is a binary variable called **FBS** (Fasting Blood Sugar) and it shows whether a patient had a fasting blood sugar level above 120 mg/dl (1) or not (0). The sixth variable is **Restecg** (resting electrocardiographic results) and it is a discrete variable showing three different levels of results, 0, 1, or 2. The seventh variable measuring the highest heart rate achieved is called **Thalach**, and it is a continuous variable. The eighth variable is called **Exang** and it is a binary variable showing if the patient had an exercise-induced angina (1) or not (0). The ninth variable called **Oldpeak** is a continuous variable showing an ST depression induced by exercise relative to rest. The tenth variable is a discrete variable called **Slope** with values 0, 1, and 2 showing the slope of the peak exercise ST segment. The eleventh variable called **Ca** is a discrete variable with values of 0, 1, 2, and 3 that shows the number of vessels colored by fluoroscopy. The twelfth variable is a discrete variable called **Thal** and it shows the different levels of a blood disorder called Thalassemia. At 0, the value is null and it is removed from the data set. At 1, there is a fixed defect and there is no blood flow in some part of the heart. At 2, there is normal blood flow. At 3, it indicates a reversible defect, and blood flow is observed but not normal. The thirteenth variable is called **Trestbps** which is a continuous variable measuring the patient's resting blood pressure. Finally, the final variable is called **Target** which is a binary variable indicating if the patient had heart disease (1) or not (0). This variable will also be our response variable when training our models.

Data Exploration (Ariana Quilantan & Angel Rodriguez)

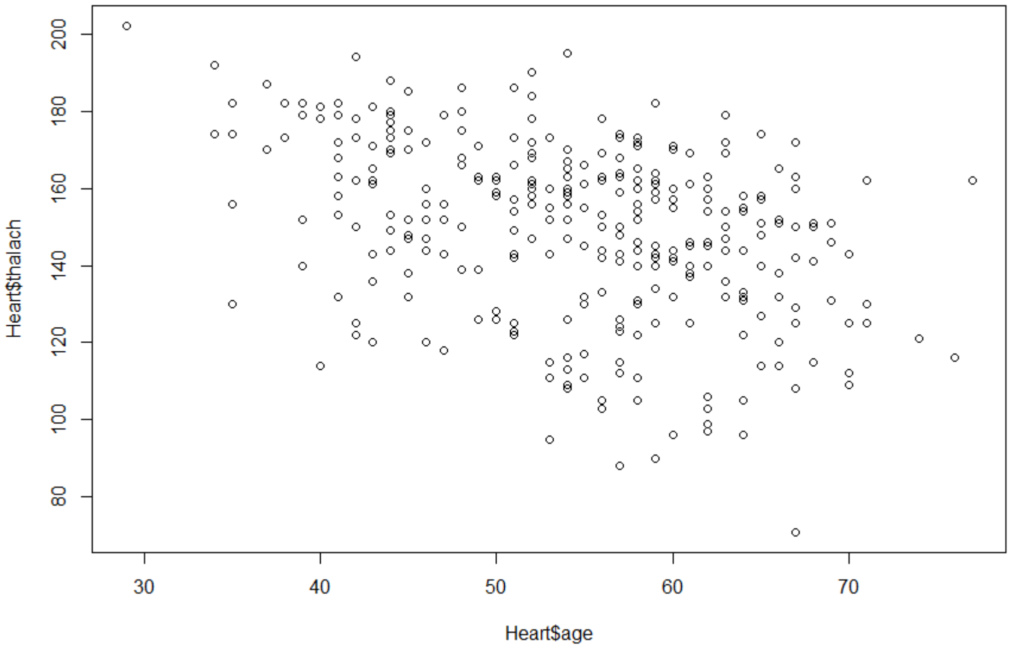
As we began to explore the data, we first observed that the average age of the patients was 54.43 years of age. 69.56% of these patients were male and 42.08% of these were diagnosed with heart disease. The remaining 30.44% were female and 27.56% of them were diagnosed with heart disease. As we can see from the figure below, a larger proportion of men was diagnosed with heart disease when compared to women.



When looking at Fasting Blood Sugar, we observe that 14.93% of all patients had values above 120 mg/dl and the average cholesterol reading was 246 mg/dl. When running a summary of the Thalach variable, we see that the minimum value for the highest heart rate achieved was only 71 bps while the maximum value was 202 bps. The average highest heart rate was 149 bps while the median was only 152 bps as we can see from the box plot below.



We also created a scatter plot between Age and Thalac (highest heart rate achieved) and observed how there is a small negative relationship between the two. As Age increases, the highest heart rate achieved appears to decrease. This is shown in the plot below.



We further explored the remaining variables, but none showed any correlation with others. There was also nothing out of the ordinary that we decided to include in this report. We therefore concluded that every variable could be left in the data set and be used in creating our models.

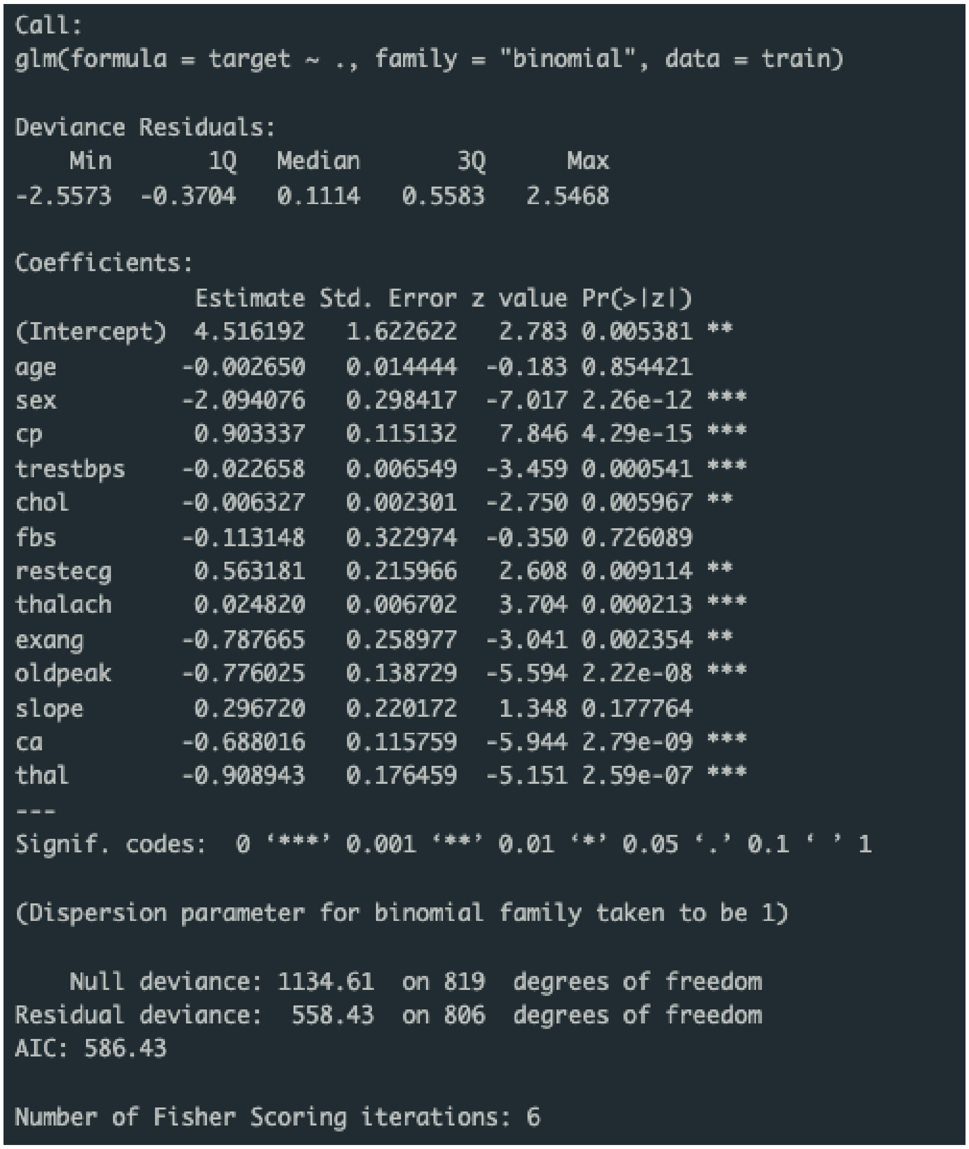
Logistic Regression (Alexander Hernandez)

We chose Logistic Regression because it is a good machine learning model to use in predicting a qualitative binary response variable like in our project. Logistic Regression takes an already good model, Linear Regression, and modifies it for classification needs by taking the logit of the linear equation. This creates a probability of being in one class or the other. The values range from 0 to 1. For this model we will determine any probability value greater than 0.50 as the patient having heart disease, otherwise, it means the patient does not have heart disease.

Model Formula

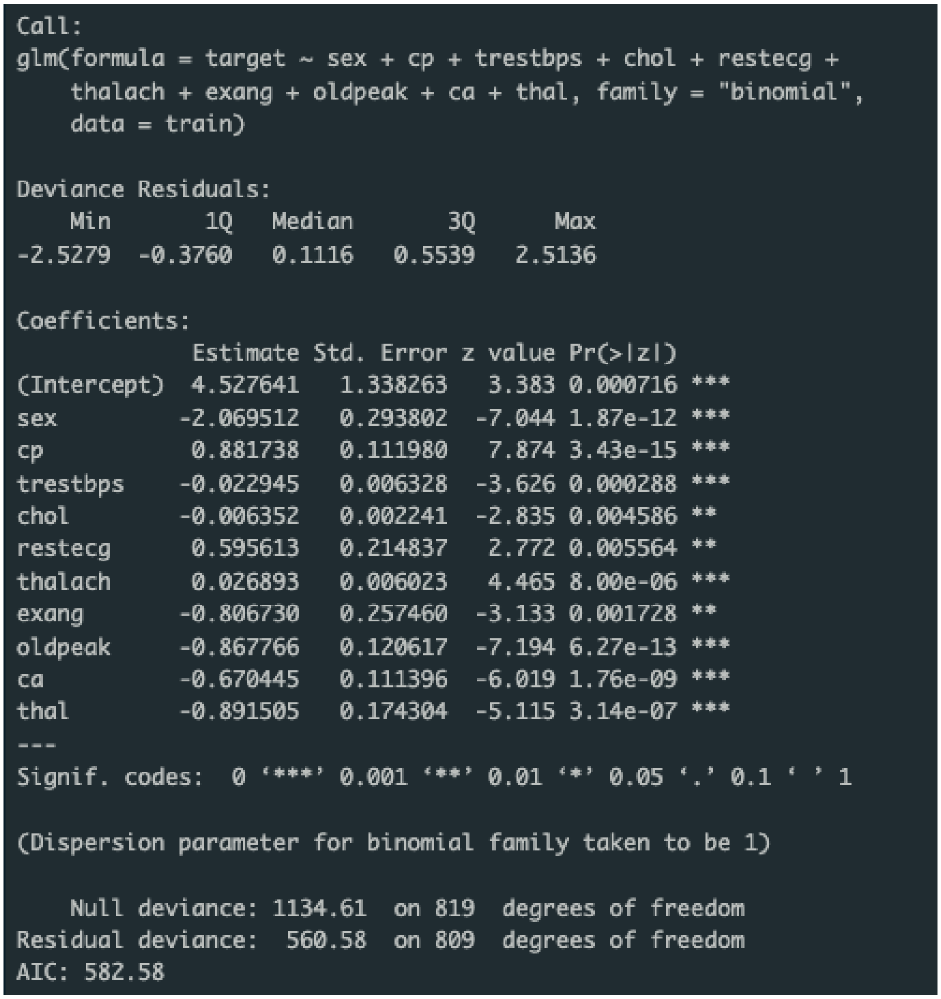
$$P(target\ =\ 1)\ =\ \frac{1}{1+e^{-(b_0+\ b_1*age+\ b_2*sex+b_3*cp+b_4*trestbps+b_5*chol\ +\ b_6*fbs\ +b_7*restecg\ +b_8*thalach+b_9*exang\ +b_{10}*oldpeak+b_{11}*slope+b_{12}*ca+b_{13}*thal)}}$$

Initially, we will consider all of the predictors and create a logistic regression model to see which specific variables R considers significant/insignificant. We will use an 80/20 training/test split in order to have data to test our model. Hence, using the glm() function we will train the model and check the summary.



```
Call:
glm(formula = target ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5573  -0.3704   0.1114   0.5583   2.5468

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.516192   1.622622   2.783 0.005381 **
age         -0.002650   0.014444  -0.183 0.854421
sex         -2.094076   0.298417  -7.017 2.26e-12 ***
cp           0.903337   0.115132   7.846 4.29e-15 ***
trestbps    -0.022658   0.006549  -3.459 0.000541 ***
chol        -0.006327   0.002301  -2.750 0.005967 **
fbs         -0.113148   0.322974  -0.350 0.726089
restecg      0.563181   0.215966   2.608 0.009114 **
thalach      0.024820   0.006702   3.704 0.000213 ***
exang       -0.787665   0.258977  -3.041 0.002354 **
oldpeak     -0.776025   0.138729  -5.594 2.22e-08 ***
slope        0.296720   0.220172   1.348 0.177764
ca          -0.688016   0.115759  -5.944 2.79e-09 ***
thal        -0.908943   0.176459  -5.151 2.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1134.61  on 819  degrees of freedom
Residual deviance:  558.43  on 806  degrees of freedom
AIC: 586.43

Number of Fisher Scoring iterations: 6
```

Here we see that most of the predictors are significant in predicting whether or not someone has heart disease. Specifically, the variables not included are **age**, **fbs,** and **slope**. It's also worth noting that the AIC for the entire model is 586.43. We will attempt to find the best possible subset of predictors that will reduce the AIC. To do this, we'll use the step() function with the backward direction so that we start with the full set of predictors and gradually remove the variables not needed.



```
Call:
glm(formula = target ~ sex + cp + trestbps + chol + restecg +
    thalach + exang + oldpeak + ca + thal, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5279  -0.3760   0.1116   0.5539   2.5136

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.527641   1.338263   3.383 0.000716 ***
sex         -2.069512   0.293802  -7.044 1.87e-12 ***
cp           0.881738   0.111980   7.874 3.43e-15 ***
trestbps    -0.022945   0.006328  -3.626 0.000288 ***
chol        -0.006352   0.002241  -2.835 0.004586 **
restecg      0.595613   0.214837   2.772 0.005564 **
thalach      0.026893   0.006023   4.465 8.00e-06 ***
exang       -0.806730   0.257460  -3.133 0.001728 **
oldpeak     -0.867766   0.120617  -7.194 6.27e-13 ***
ca          -0.670445   0.111396  -6.019 1.76e-09 ***
thal        -0.891505   0.174304  -5.115 3.14e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1134.61  on 819  degrees of freedom
Residual deviance:  560.58  on 809  degrees of freedom
AIC: 582.58
```

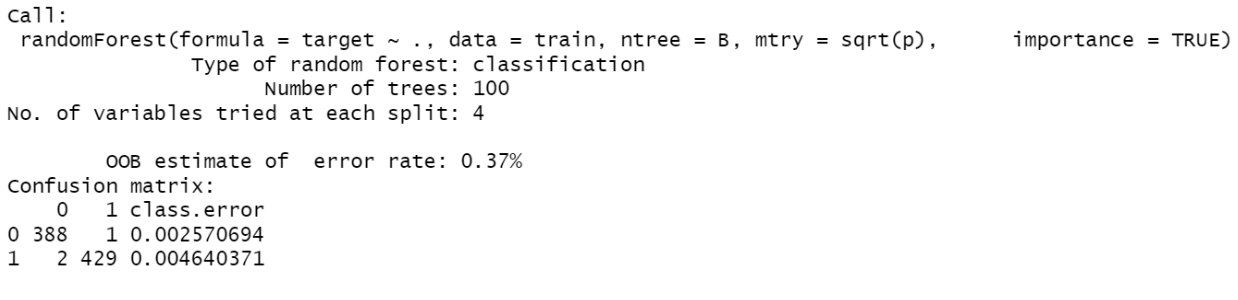
Just like we predicted based on the summary, the step function removed those variables that were least significant. We also were able to reduce the AIC slightly from 586.43 to 582.58. Hence the final logistic regression model will have the form of:

$$P(target\ =\ 1)\ =\frac{1}{1+e^{-(b_0+\ b_1*sex+b_2*cp+b_3*trestbps+b_4*chol\ +b_5*restecg\ +b_6*thalach+b_7*exang\ +b_8*oldpeak+b_9*ca+b_9*thal)}}$$

Now that we have the best subset of predictors we will create 10 different 80/20 training/test splits and then record the prediction errors by creating a confusion matrix of the predicted and observed targets and calculate the test error by calculating (FP + FN)/(TP + TN + FP + FN). This results in test errors being [0.1658537, 0.1804878, 0.1365854, 0.1756098, 0.1317073, 0.1707317, 0.1707317, 0.1414634, 0.1658537, 0.1609756]. The average test error is 16%. Performing the same for specificity and sensitivity we get values of 0.7950718 and 0.8810381 respectively. This means that the model detects 88% of patients with the disease, but around 12% of patients go undetected.

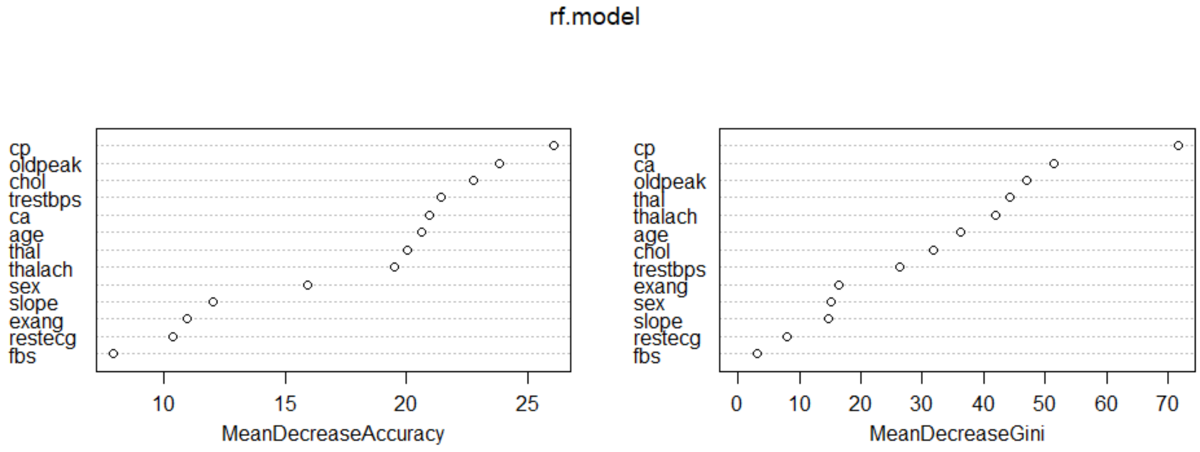
Random Forest (Salik Faisal)

Similar to a logistic model, a random forest model also captures non-linear relationships between the variables. A random forest model is also particularly useful because it allows us to see which variables or predictors are more predictive of heart disease in a patient. The MeanDecreaseAccuracy plot and MeanDecreaseGini plots, as shown below, help us visualize which variables are most important in predicting heart disease and which variables would most impact the accuracy of the predictions when removed from the model. We initially set B = 100, which indicates 100 trees, and divided the data into 80% training and 20% testing. Because the model uses classification trees, **mtry** is set equal to the square root of the total number of predictors. Using the training data to develop the random forest model, we get the following results:



```
Call:
 randomForest(formula = target ~ ., data = train, ntree = B, mtry = sqrt(p),      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 4

        OOB estimate of  error rate: 0.37%
Confusion matrix:
    0   1 class.error
0 388   1 0.002570694
1   2 429 0.004640371
```

The OOB estimate of the error rate or misclassification rate is 0.37%. This is a drastic improvement compared to the logistic model. Because the error rate is at an ideal result, we do not need to remove any variables from the original random forest model. This also means that in the random first model, ALL of the variables are predictive of whether a patient has heart disease or not.



After running the random forest model on 10 different 80%-20% training-testing splits, we get an average misclassification or error rate of **0.73%**. The error rate indicates that a random forest model is not only superior to a logistic model but also very successful in predicting heart disease in a patient overall. The average sensitivity rate across all 10 iterations of the model is **99.08%**. This means that the random forest is able to accurately detect heart disease in a patient over 99% of the time. The specificity rate is **99.46%**. This is the rate of patients who do not have heart disease predicted correctly to not have it.

Model Formula

$target \sim cp + oldpeak + chol + trestbps + ca + age + thal + thalach + sex + slope + exang + restecg + fbs$

Conclusion (Angel Rodriguez)

After having fitted both a logistic regression and a random forest model to the data, we were able to conclude that the most important variables for predicting whether a patient has heart disease or not when running a logistic regression are sex, chest pain (cp), resting blood pressure (trestbps), cholesterol (chol), resting electrocardiographic results (restecg), highest heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), number of vessels colored by fluoroscopy (ca), and the level of thalassemia (thal). For random forests, the variables are cp, oldpeak, chol, trestbps, ca, age, thal, thalach, and sex. Both models include sex, cp, trestbps, chol, oldpeak, ca, and thal. Random forests include two more: thalach and age. According to the CDC, many of the causes of heart disease have more to do with dietary and physical activity. Other conditions such as obesity and diabetes are leading causes of heart disease. Those two are not observed or measured in our data but it is possible to study the relationship between what we have observed and whether or not that patient also has heart disease. Age is also considered to be another factor in being diagnosed with heart disease. There is the possibility that our dataset measures certain variables that may be predictors of other conditions which then lead to heart disease. It is safe to say both of our models calculated that similar variables were important in predicting heart disease. However, we are also interested in finding out which model was best.

In order to compare both models and their accuracy we went based on the misclassification/error rate. For the logistic regression model, we calculated an error rate of 16%. For the random forest model, we calculated a drastically lower error rate of .73%. This did not come as a surprise since it is known that random forests are extremely accurate when the number of predictor variables increases.

Heart disease affects millions of people annually and discovering better methods of predicting heart disease could potentially help doctors administer treatment earlier in the process, and could help patients identify certain activities that may be causing heart disease. The first variable marked as a variable of importance was the level of chest pain (cp) experienced by a patient. Further analysis of this correlation between chest pain and heart disease is needed.

Bibliography

1. Lapp, D. (2019, June 6). *Heart disease dataset*. Kaggle. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download
2. National Center for Health Statistics. Multiple Cause of Death 2018–2021 on CDC WONDER Database. Accessed February 2, 2023.

3. National Center for Health Statistics. Percentage of coronary heart disease for adults aged 18 and over, United States, 2019—2021. National Health Interview Survey. Accessed February 17, 2023.