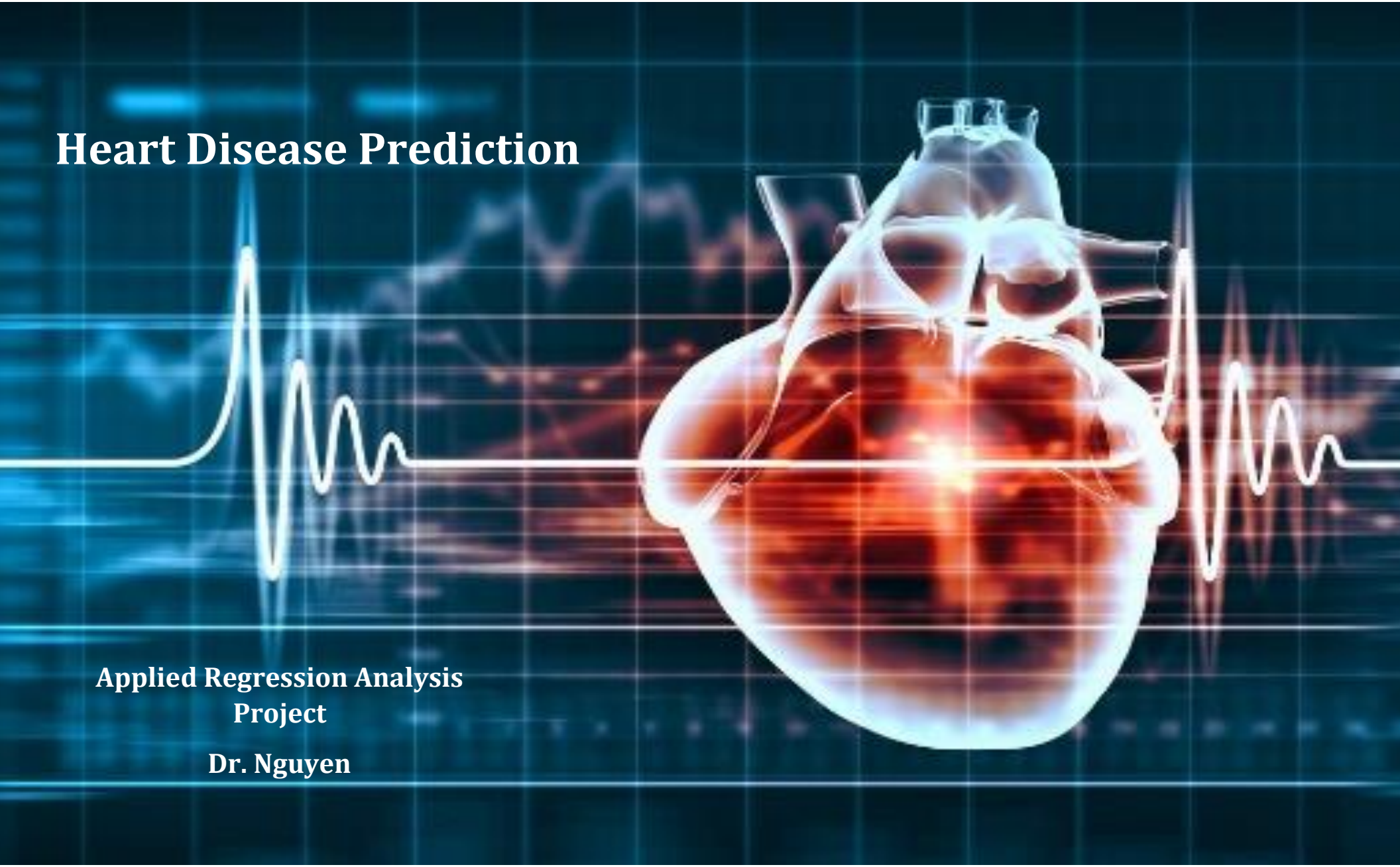
Heart Disease Prediction



Applied Regression Analysis Project

Dr. Nguyen

Team Members: Karen Nogueira, Francisco Ortiz, Mudiha Wazirali

ABSTRACT

The Cleveland Heart Disease Data found in the UCI machine learning repository consists of 14 variables measured on 303 individuals. Each person has been classified as 1 indicating presence of Heart Disease, or 0 indicating absence of it. This variable has been named as the “target”.

Because of the binary classification problem found in this dataset, we aim to explore a logistic regression model capable to predict the target value based on the variables which provide statistically significance to that prediction.

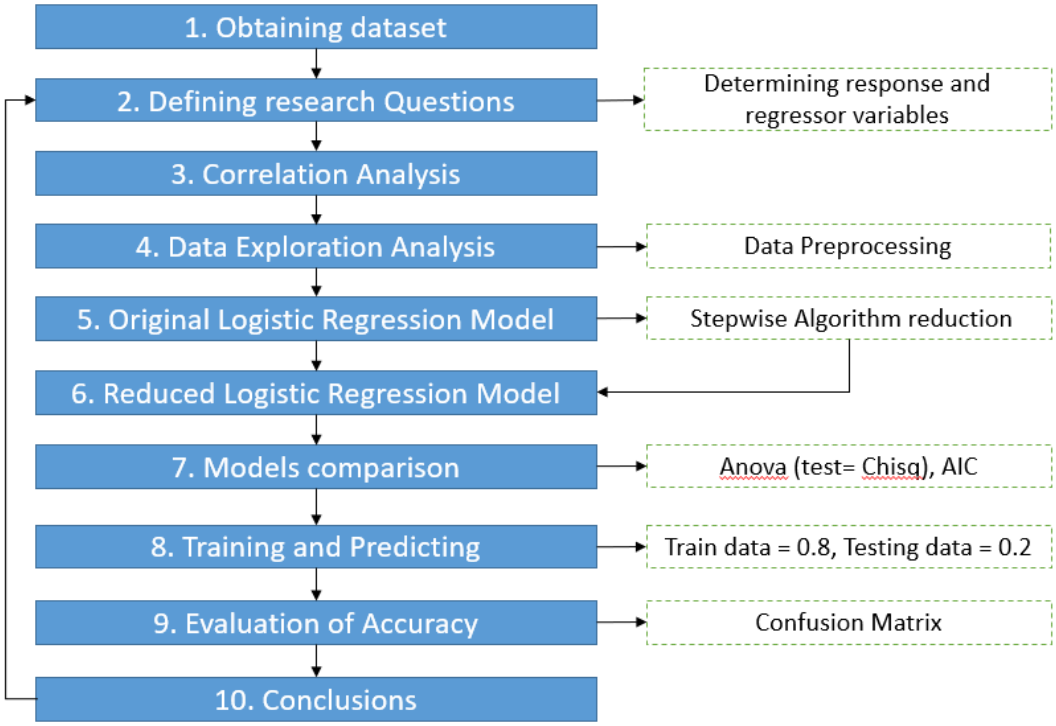
The results indicate that 8 regressor variables from the original13 allow for a prediction accuracy of 90% based on a logistic regression model.

OBJECTIVES

- To understand the relationship between the regressor variables and the presence or absence of heart disease based on a correlation analysis and a general data exploratory analysis.
- To build a parsimonious model capable to predict the target variable by selecting the regressors that explain most of the variability in the dataset.
- To develop the best regression model that predicts the probability to either have a heart disease or not.
- To identify the most important variables that predict a heart disease.

DATA ANALYSIS FLOW

In order to obtain results and answer our objective questions, we defined the next data analysis flow:



1. DATASET

The dataset, which the main source is found in the Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease), has been obtained from Kaggle following the next link: https://www.kaggle.com/ronitf/heart-disease-uci.
Specifically, we will work with the Cleveland Heart Disease Data which consists of 14 variables measured on 303 individuals.

The 14 variables included in the dataset are described in the next section.

VARIABLES DEFINITION

Variable Name	Class	Description
age	Numerical	The person's age in years
sex	Factor	The person's genre. 1 = male, 0 = female
cp	Factor	The chest pain type experienced; • 0 = typical angina • 1 = atypical angina • 2 = non-anginal pain • 3 = asymptomatic
trestbps	Numerical	The person's resting blood pressure (mm Hg on admission to the hospital)
chol	Numerical	The person's cholesterol serum measurement in mg/dl
fbs	Factor	The person's fasting blood sugar > 120 mg/dl; 1 = true, 0 = false
restecg	Factor	Resting electrocardiographic measurement; • 0 = normal • 1 = having ST-T wave abnormality • 2 = showing probable or definite left ventricular hypertrophy
thalach	Numerical	The person's maximum heart rate achieved
exang	Factor	Exercise induced angina; 1 = yes, 0 = no
oldpeak	Numerical	ST depression induced by exercise relative to rest ('ST' relates to positions on the EKG plot)
slope	Factor	The slope of the peak exercise ST segment; 3 levels. • 0 = upsloping • 1 = flat • 2 = downsloping
ca	Factor	The number of major vessels colored by fluoroscopy: 0-4 vessels
thal	Factor	Thallium stress test level. • 0 -1 = normal • 2 = fixed defect • 3 = reversable defect
target	Factor	Heart disease. 0 = no, 1 = yes

Initial Data Structure

```
data.frame':   303 obs. of  14 variables:
$ ï..age  : int  63 37 41 56 57 57 56 44 52 57 ...
$ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
$ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
$ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
$ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
$ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
$ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
$ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
$ exang   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
$ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
$ slope   : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
$ ca      : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
$ thal    : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
$ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

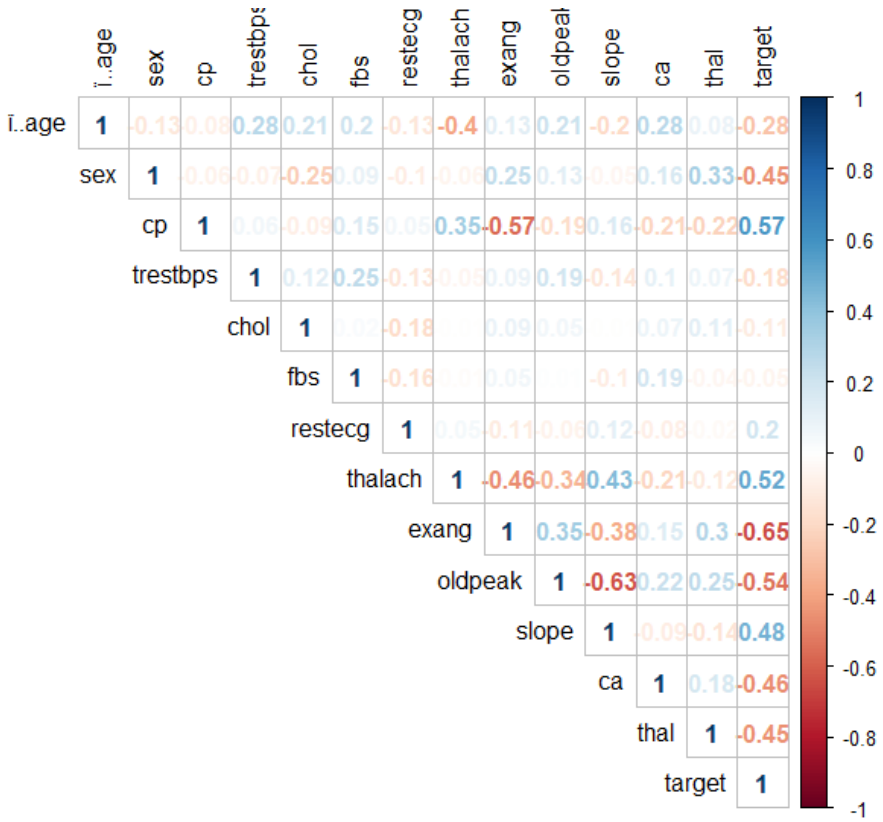
2. RESEARCH QUESTIONS

Based on the objectives mentioned before, we want to answer the next questions:

- What are the relationships between the regressors and the target variable?
- What is the best model to predict a heart disease state?

3. CORRELATION ANALYSIS

Using a heterogeneous data correlation analysis, the results are shown in the next graphic:

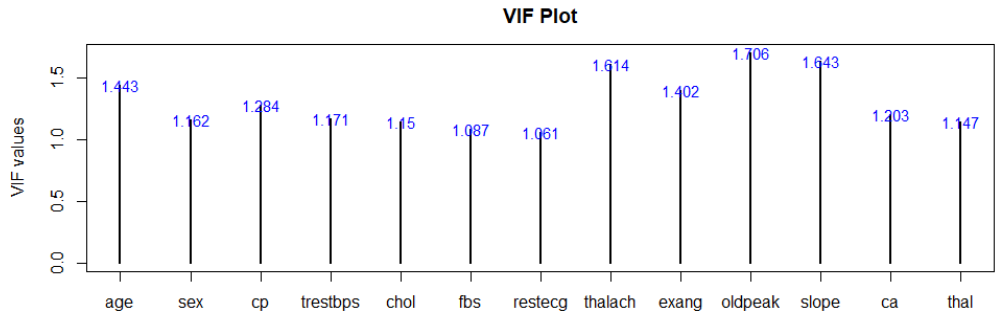


This correlation graphic highlights the correlations between regressors and between target and regressors. Positive correlations are shown in blue, and negative correlations are in red. In addition, color intensity is directly related to the grade of the correlation coefficients. Thus, we have:

- There is a negative correlation of -0.45 between the target and the variable sex.
- There is a negative correlation of -0.65 between the target and the variable exang.
- There is a negative correlation of -0.54 between the target and the variable oldpeak.
- There is a negative correlation of -0.46 and -0.45 between the target and the variables ca and thal respectively.
- There is a positve correlation of 0.57 between the target and the variable cp.
- There is a positve correlation of 0.52 between the target and the variable thalach.
- There is a positve correlation of 0.48 between the target and the variable slope.
- There is a negative correlation of -0.57 between the variables exang and cp.
- There is a negative correlation of -0.63 between the variables oldpeak and slope.

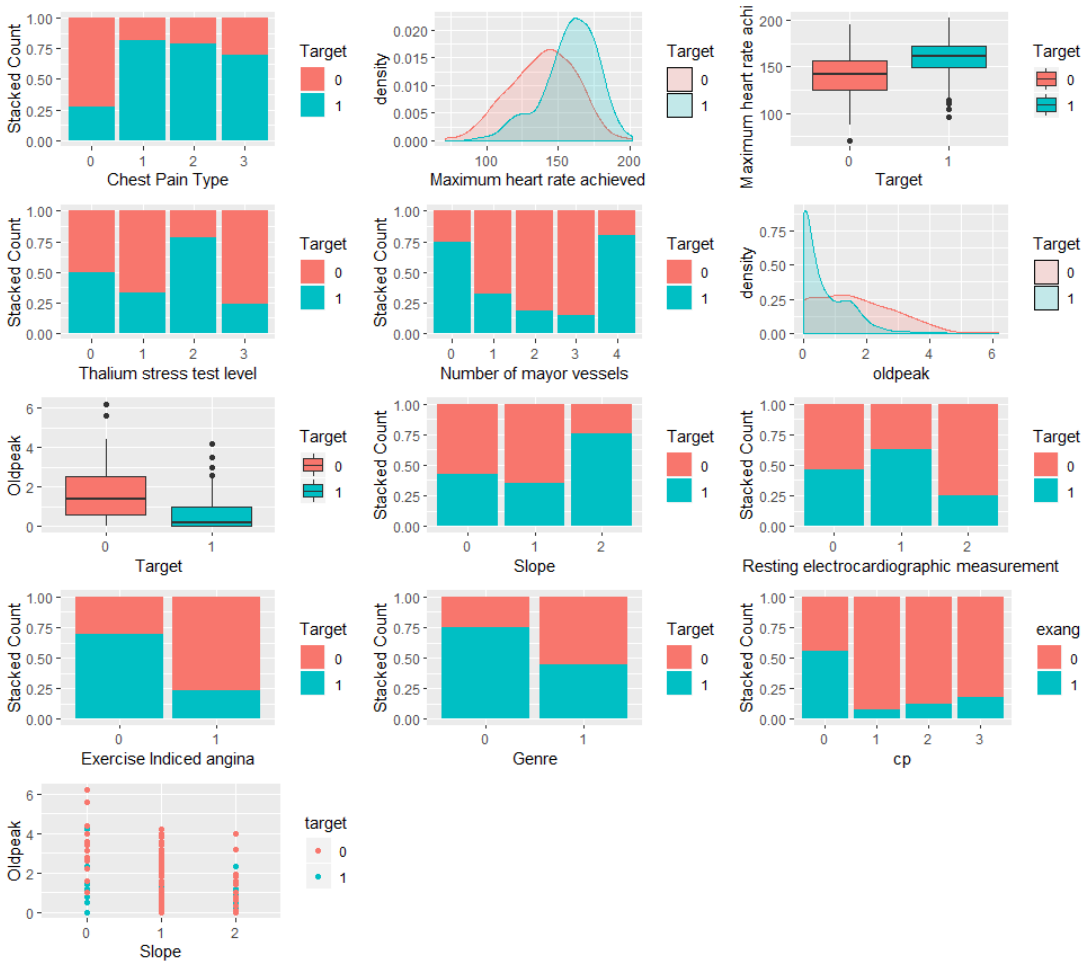
Eight variables show a moderate positive or negative correlation related to the target. These variables are sex, cp, thalach, exang, oldpeak, slope, ca, and thal. These may be the features that play a more significant role in driving the prediction of the target.

The next VIF plot indicates that there is not a problem of multicollinearity, and so we don't delete any variable based on this.



4. DATA EXPLORATION ANALYSIS

According to the correlation results, we identified the variables to perform the data exploration analysis to understand how they are related, by visualizing their relationships and reviewing the available literature. The next figure shows the relationships.



Next are the relations that resulted by plotting the variables:

- **Chest pain and target:**
  the cp type 1 = “Atypical angina” has the biggest incidence of heart disease compared to the others. However, the graph indicates that cp type 2 and 3 have an important relationship to heart disease.

Angina is a type of chest pain caused by reduced blood flow to the heart. is often described as squeezing, pressure, heaviness, tightness or pain in the chest. During stable angina, episodes of chest discomfort are usually predictable. They can occur during exertion (such as running) or during mental or emotional stress. Normally, the chest discomfort is relieved with rest, use of nitroglycerin, or both. In unstable angina, chest pain can occur at any time—often while a person is resting. The discomfort may be more severe and last longer than in typical angina. The most common cause is reduced blood flow to the heart muscle because the coronary arteries are narrowed by fatty buildups.

There is an atypical case of angina known as Prinzmetal angina. While in general the outlook of patients with Prinzmental angina is quite good, this condition can cause serious problems. It can trigger dangerous and potentially fatal cardiac arrhythmias, especially ventricular fibrillation. And while heart attacks are uncommon with Prinzmetal angina, they indeed can occur, producing permanent damage to the heart muscle.

Prinzmetal angina occurs when an area within one of the major coronary arteries suddenly goes into spasm, temporarily shutting off blood flow to the heart muscle supplied by that artery. During these episodes, the electrocardiogram (ECG) shows dramatic elevations of the "ST segment" — the same ECG changes commonly seen with heart attacks.

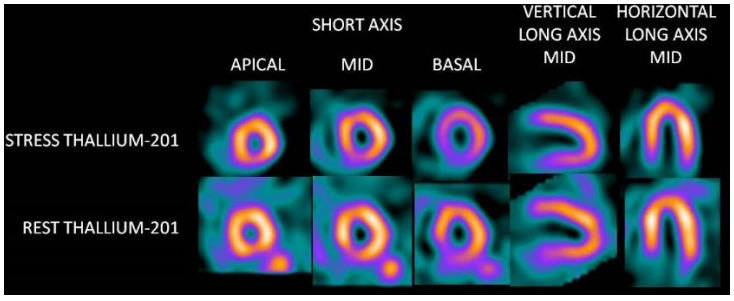
We can see that the type of angina is directly related to the fact to present a heart disease.

- **Thalach and target:**
  the heart disease count is present when the thalach is over 150. The median of this variable is 153

A normal heart rate is usually stated as 60 to 100 beats per minute. Slower than 60 is bradycardia ("slow heart"); faster than 100 is tachycardia ("fast heart").

- **Thal and target:**
  The stress level of 2 (fixed defect) indicates the mayor incidence to have heart disease.

A thallium stress test is a nuclear imaging test that shows how well blood flows into the heart while the patient is exercising or at rest (as in the next figure). This test is also called a cardiac or nuclear stress test. Abnormal results may indicate: Reduced blood flow to part of the heart caused by narrowing or blockage of one or more arteries that supply your heart muscle; Scarring of the heart muscle due to a previous heart attack; heart disease; a too-large heart, indicating other heart complications.



- **ca and target:**
  0 and 4 vessels indicate a mayor incidence to have heart disease

Five great vessels enter and leave the heart: the superior and inferior vena cava, the pulmonary artery, the pulmonary vein, and the aorta. Doctors use angiography to diagnose and treat blood vessel diseases and conditions. Angiography exams produce pictures of major blood vessels throughout the body. In some cases, contrast material is used.

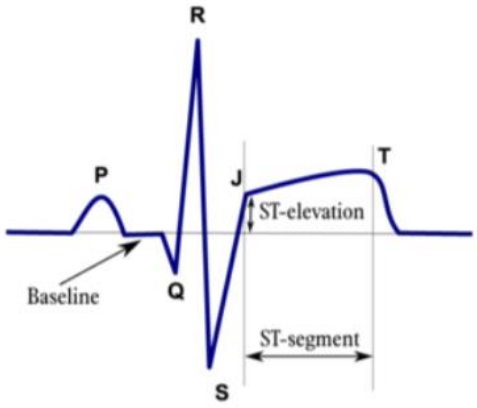
- **Oldpeak and target:**
  Lower Oldpeak (or depression) indicates heart disease

- **Slope and target:**
  slope of Level 2 = downsloping indicates heart disease.

- **Slope and oldpeak:**
  there is more incidence of heart disease for patients with a slope of 2 and low oldpeak

An electrocardiogram — abbreviated as EKG or ECG — is a test that measures the electrical activity of the heartbeat. With each beat, an electrical impulse (or “wave”) travels through the heart. This wave causes the muscle to squeeze and pump blood from the heart. A normal heartbeat on ECG will show the timing of the top and lower chambers.

For a better understanding of variables describing the EKG results, we can see in the next figure that a normal ST segment has a slight upward concavity. Flat, downsloping or depressed ST segments may indicate coronary ischemia. ST elevation may indicate transmural myocardial infarction.



The right and left atria or upper chambers make the first wave called a “P wave" — following a flat line when the electrical impulse goes to the bottom chambers. The right and left bottom chambers or ventricles make the next wave called a “QRS complex." The final wave or “T wave” represents electrical recovery or return to a resting state for the ventricles.

The variables that describe ECG: Restecg, Oldpeak, and Slope.

- **Restecg and target:**
  Restecg of 1 indicates having ST-T wave abnormality. The majority of observations with a heart disease had this measure. Interestingly, some of the patients with a normal level of 0 had a heart disease event.

- **Exang and target:**
  The level 0=no exercise angina (exang) indicate a count with patients with heart disease compared to level 1 = yes

- **Exang and cp:**
  For a typical angina there is more incidence to have an exercise induced angina.

- **Gender and target:**
  women had more incidence of heart disease compared to men (in proportion to each genre).

It is known that gender plays a role in the symptoms, treatments and outcomes of coronary artery disease (CAD). According to the Texas Heart Institute, cardiovascular diseases affect more women than men and are responsible for more than 40% of all deaths in American women.

A simple linear regression model indicates that the features that had a moderate correlation with the target variable also have significant p-values. This points towards a possible reduction in the needed number of variables when shaping an optimal model for prediction. This could result in an improvement of the R-squared value which is 0.5175 considering all the regressors in the model.

```
Call:
lm(formula = target ~ ., data = heart2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94748 -0.21270  0.06608  0.25022  0.93509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8288987  0.2929344   2.830 0.004987 **
age         -0.0008204  0.0026962  -0.304 0.761129
sex         -0.1959956  0.0471429  -4.157 4.24e-05 ***
cp           0.1127034  0.0223816   5.036 8.40e-07 ***
trestbps    -0.0019910  0.0012573  -1.583 0.114407
chol        -0.0003535  0.0004217  -0.838 0.402545
fbs          0.0173736  0.0596669   0.291 0.771125
restecg      0.0498480  0.0399228   1.249 0.212819
thalach      0.0030193  0.0011304   2.671 0.007988 **
exang       -0.1440459  0.0513689  -2.804 0.005387 **
oldpeak     -0.0587887  0.0229269  -2.564 0.010847 *
slope        0.0789788  0.0423896   1.863 0.063453 .
ca          -0.1006022  0.0218565  -4.603 6.25e-06 ***
thal        -0.1190392  0.0356550  -3.339 0.000952 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3542 on 289 degrees of freedom
Multiple R-squared:  0.5175,    Adjusted R-squared:  0.4958
F-statistic: 23.85 on 13 and 289 DF,  p-value: < 2.2e-16
```

LOGISTIC REGRESSION MODEL

5. ORIGINAL LOGISTIC REGRESSION MODEL

In this model all the variables are included in the regression equation resulting in an AIC of **225.63**

Applying a Stepwise Algorithm in direction backward, the lowest AIC of 219.79 is achieved by dropping the variables age, fbs, chol, restecg, and thalach from the original dataset. Most of these variables were suggested by the imndiag() function as non-significant when testing fo multicollinearity.

```
            VIF    TOL      Wi      Fi Leamer    CVIF Klein   IND1   IND2
i..age   1.4435 0.6928 10.7173 11.7319 0.8323 -2.2366     0 0.0287 1.4009
sex      1.1619 0.8607  3.9118  4.2821 0.9277 -1.8002     0 0.0356 0.6353
cp       1.2845 0.7785  6.8743  7.5251 0.8823 -1.9902     0 0.0322 1.0098
trestbps 1.1706 0.8543  4.1226  4.5129 0.9243 -1.8137     0 0.0353 0.6645
chol     1.1502 0.8694  3.6292  3.9728 0.9324 -1.7821     0 0.0360 0.5954
fbs      1.0874 0.9196  2.1117  2.3116 0.9590 -1.6848     0 0.0381 0.3664
restecg  1.0610 0.9425  1.4741  1.6137 0.9708 -1.6439     0 0.0390 0.2621
thalach  1.6137 0.6197 14.8317 16.2359 0.7872 -2.5003     0 0.0256 1.7342
exang    1.4020 0.7133  9.7150 10.6348 0.8446 -2.1723     0 0.0295 1.3074
oldpeak  1.7059 0.5862 17.0582 18.6731 0.7656 -2.6431     0 0.0243 1.8868
slope    1.6426 0.6088 15.5294 16.9995 0.7803 -2.5451     0 0.0252 1.7838
ca       1.2026 0.8316  4.8954  5.3589 0.9119 -1.8633     0 0.0344 0.7681
thal     1.1473 0.8716  3.5592  3.8962 0.9336 -1.7776     0 0.0361 0.5853

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

i..age , trestbps , chol , fbs , restecg , slope , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.5175
```

6. REDUCED LOGISTIC REGRESSION MODEL

According to the variable reduction procedure, we generated a dataset modified which contains 8 variables: exang, trestbps (resting blood pressure), oldpeak, thal, sex, slope, cp, ca.

The result from the reduced logistical regression model results on an AIC of **219.79**, lower than I the original model.

7. MODELS COMPARISON

Both the original model and reduced model are shown in the next figure.

Original model (AIC = 225.63)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9459  -0.2738   0.1012   0.4515   3.1248

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.179045   3.705420   0.048 0.961461
i..age       0.027819   0.025428   1.094 0.273938
sex1        -1.862297   0.570844  -3.262 0.001105 **
cp1          0.864708   0.578000   1.496 0.134645
cp2          2.003186   0.529356   3.784 0.000154 ***
cp3          2.417107   0.719242   3.361 0.000778 ***
trestbps    -0.026162   0.011943  -2.191 0.028481 *
chol        -0.004291   0.004245  -1.011 0.312053
fbs1         0.445666   0.587977   0.758 0.448472
restecg1     0.460582   0.399615   1.153 0.249089
restecg2    -0.714204   2.768873  -0.258 0.796453
thalach      0.020055   0.011859   1.691 0.090820 .
exang1      -0.779111   0.451839  -1.724 0.084652 .
oldpeak     -0.397174   0.242346  -1.639 0.101239
slope1      -0.775084   0.880495  -0.880 0.378707
slope2       0.689965   0.947657   0.728 0.466568
ca1         -2.342301   0.527416  -4.441 8.95e-06 ***
ca2         -3.483178   0.811640  -4.292 1.77e-05 ***
ca3         -2.247144   0.937629  -2.397 0.016547 *
ca4          1.267961   1.720014   0.737 0.461013
thal1        2.637558   2.684285   0.983 0.325808
thal2        2.367747   2.596159   0.912 0.361759
thal3        0.915115   2.600380   0.352 0.724901
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 179.63  on 280  degrees of freedom
AIC: 225.63

Number of Fisher Scoring iterations: 6
```

Reduced model (AIC = 219.79)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0264  -0.3305   0.1082   0.4577   2.9873

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.30454    4.33809   0.762 0.446210
exang1      -0.85234    0.43663  -1.952 0.050925 .
slope1      -0.90675    0.82751  -1.096 0.273189
slope2       0.70078    0.89682   0.781 0.434570
trestbps    -0.02211    0.01093  -2.023 0.043051 *
oldpeak     -0.47970    0.23137  -2.073 0.038142 *
thal1        2.62410    4.07723   0.644 0.519836
thal2        2.36301    4.01217   0.589 0.555888
thal3        0.91673    4.01441   0.228 0.819366
sex1        -1.63154    0.52712  -3.095 0.001967 **
cp1          1.03058    0.56475   1.825 0.068023 .
cp2          2.22015    0.51711   4.293 1.76e-05 ***
cp3          2.55944    0.70231   3.644 0.000268 ***
ca1         -2.35513    0.49869  -4.723 2.33e-06 ***
ca2         -3.10939    0.75361  -4.126 3.69e-05 ***
ca3         -2.26756    0.90067  -2.518 0.011815 *
ca4          1.23217    1.62177   0.760 0.447393
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 185.79  on 286  degrees of freedom
AIC: 219.79

Number of Fisher Scoring iterations: 6
```

The test for significance has been applied by performing ANOVA test. The p-values are calculated using the chi-squared distribution, but like the parametric alternative they indicate whether each of the predictors has a significant effect on the probability of achieving an indicator value of 1.

The next figure shows the anova test for both the original and reduced models.

In the original model we see that the variables chol, fbs, and restecg are not significant for the model, whereas for the reduced model all the variables are significant at different levels of significance.

Original model

```
> lr.anova <- anova(heart_mod1, test="Chisq")
> lr.anova
Analysis of Deviance Table

Model: binomial, link: logit

Response: target

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       302     417.64
i..age    1   15.777       301     401.86 7.128e-05 ***
sex       1   31.287       300     370.57 2.225e-08 ***
cp        3   77.654       297     292.92 < 2.2e-16 ***
trestbps  1    4.925       296     287.99  0.026464 *
chol      1    2.376       295     285.62  0.123181
fbs       1    0.024       294     285.59  0.875956
restecg   2    3.212       292     282.38  0.200670
thalach   1   19.406       291     262.98 1.057e-05 ***
exang     1    5.612       290     257.36  0.017841 *
oldpeak   1   17.196       289     240.17 3.371e-05 ***
slope     2    5.440       287     234.73  0.065865 .
ca        4   41.394       283     193.33 2.228e-08 ***
thal      3   13.703       280     179.63  0.003339 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reduced model

```
> lr.anova_red <- anova(heart_mod2, test="Chisq")
> lr.anova_red
Analysis of Deviance Table

Model: binomial, link: logit

Response: target

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       302     417.64
exang     1   59.735       301     357.90 1.085e-14 ***
slope     2   29.121       299     328.78 4.747e-07 ***
trestbps  1    4.346       298     324.44   0.03710 *
oldpeak   1   20.581       297     303.85 5.716e-06 ***
thal      3   37.759       294     266.10 3.179e-08 ***
sex       1    5.892       293     260.20   0.01521 *
cp        3   29.857       290     230.35 1.479e-06 ***
ca        4   44.562       286     185.79 4.903e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, performing an ANOVA test comparing both the original and reduced models, the P-value based on chi-squared distribution indicates that we fail in reject the null Hypothesis which estates that the two models are equal. What this means is that the reduced model is better that the original containing all the regressor variables.

```
Analysis of Deviance Table

Model 1: target ~ i..age + sex + cp + trestbps + chol + fbs + restecg +
    thalach + exang + oldpeak + slope + ca + thal
Model 2: target ~ exang + slope + trestbps + oldpeak + thal + sex + cp +
    ca
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       280     179.63
2       286     185.78 -6  -6.1543   0.4061
```

From the reduced model, we can see that for example if all regressors are held at a fixed value, the odds of getting heart disease for males (sex=1) over the odds of getting heart disease for females is exp(-1.63154) = 0.1956 i.e. the odds are lower than the odds of the women getting heart disease. This was known already from the data exploratory analysis.

8. TRAINING AND PREDICTING BASED ON LOGISTIC REGRESSION MODEL

The distribution of the target attribute is 138 patients with positive diagnostic and 165 healthy people, indicating that there is not a problem of class imbalance, therefore, we can proceed to create a predictive model.

Prediction was performed based on both the original dataset and the reduced dataset.

For each model, a random subset of the datasets of 80% represents the training datasets, and 20% the testing datasets. For each one, there is not a problem of class imbalance.

In the reduced dataset, the logistic regression model results in the probability for an observation to be classified as 1 or 0. As an example, next is a subset of those probabilities:

```
           0         1
2  0.06652322 0.93347678
4  0.04068366 0.95931634
7  0.10056956 0.89943044
11 0.24336317 0.75663683
19 0.33146211 0.66853789
21 0.91022795 0.08977205
```

9. ACCURACY EVALUATION

Original Prediction model

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 20  1
         1  7 32

               Accuracy : 0.8667
                 95% CI : (0.7541, 0.9406)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 1.653e-07

                  Kappa : 0.7251

 Mcnemar's Test P-Value : 0.0771

            Sensitivity : 0.7407
            Specificity : 0.9697
         Pos Pred Value : 0.9524
         Neg Pred Value : 0.8205
             Prevalence : 0.4500
         Detection Rate : 0.3333
   Detection Prevalence : 0.3500
      Balanced Accuracy : 0.8552

       'Positive' Class : 0
```

Reduced Prediction model

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 22  1
         1  5 32

               Accuracy : 0.9
                 95% CI : (0.7949, 0.9624)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 4.558e-09

                  Kappa : 0.7952

 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.8148
            Specificity : 0.9697
         Pos Pred Value : 0.9565
         Neg Pred Value : 0.8649
             Prevalence : 0.4500
         Detection Rate : 0.3667
   Detection Prevalence : 0.3833
      Balanced Accuracy : 0.8923

       'Positive' Class : 0
```

For the original dataset, Logistic Regression Modeling resulted in 86% of accuracy prediction. Whereas for the reduced dataset the prediction accuracy was 90%.

Both models are more specific than sensitive, which means that they predict better false negative cases.

Because we are analyzing a human disease, a False Negative (ignoring the probability of disease there in reality there is one) is more dangerous than a False Positive indicating the case when the prediction results in 1 when in reality is 0. We see that in both models, the false negative resulted in just 1 observation.

Both models performed well based on the Area Under the Curve defined by the levels of Specificity and sensitivity. For both models the AUC= 0.945

10. CONCLUSIONS

- The variables slope, exang, trestbps, oldpeak, thal, sex, cp, and ca are the features that play a significant role in driving the prediction of the heart disease condition.
- These variables allowed for a prediction accuracy of 90% based on a logistic regression model, resulting in both a parsimonious and specific model. This is the best model built under this approach, however, further analysis by training and testing different models could be necessary to compare to the logistic regression model.

REFERENCES

https://www.texasheart.org/heart-health/heart-information-center/topics/women-and-heart-disease/

https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373