Nora Lin and Michael Morrey
IST 707 - Applied Machine Learning

Final Project: Predicting Heart Disease

Introduction

Every year in the United States, roughly 659,000 people die from heart disease, which equates to about 1 in every 4 deaths. The Centers for Disease Control and Prevention (CDC) has consistently ranked heart disease as the leading cause of death in the United States each year *(Heart Disease Facts)*. In 2020, the infamous year in which the Covid-19 pandemic took the world by storm, the heart disease death toll nearly doubled that of Covid-19 in the United States *(FASTSTATS - leading causes of death)*. On a worldwide scale, heart disease is consistently the leading killer each year, according to the World Health Organization (WHO). The WHO reports that in 2019, heart disease was responsible for 16% of the world's total deaths, and has shown the largest increase in deaths since 2000 *(The top 10 causes of death)*. These statistics make heart disease a concern for both doctors and the general public. Finding ways to predict heart disease could potentially lower these statistics for future generations.

The term "heart disease" can refer to several different types of heart conditions, the most common of which is Coronary Artery Disease (CAD). This disease occurs when arteries are narrowed or blocked, reducing blood flow to the heart. Heart disease often goes undiagnosed until a person experiences more serious symptoms, such as heart attacks. Other heart disease symptoms include chest pain, left shoulder pain, upper back pain, shortness of breath, and nausea (*Cardiovascular diseases (cvds))*. Researchers and scientists have developed various tests, such as the thallium test, which measures the amount of blood flow reaching each part of the heart. Another procedure is the fluoroscopy, which colors major blood vessels and examines the slope of exercise induced ST. All of these variables contribute to building out the multifacets of what heart disease in patients looks like.

Due to the staggering statistics behind heart disease, it is crucial to understand its risk factors. Some risk factors, such as increase in age and family medical history, are out of a person's control. Other risk factors are due to unhealthy lifestyle decisions, such as unhealthy diets, lack of physical activity, obesity, smoking, and excessive alcohol usage. Fortunately, making lifestyle changes in these areas can greatly reduce the risk of heart disease *(Types of heart disease)*. Awareness of symptoms can also give patients the ability to ask their doctors for specific tests. Through data mining and machine learning techniques, the risks for heart disease can be more accurately predicted and understood. These techniques can also be applied to other fields of medicine as well as various industries.

Analysis

About the Data

The data set is from the University of California Irvine's (UCI) machine learning repository. It is composed of data collected in Cleveland, Hungary, Switzerland and VA Long

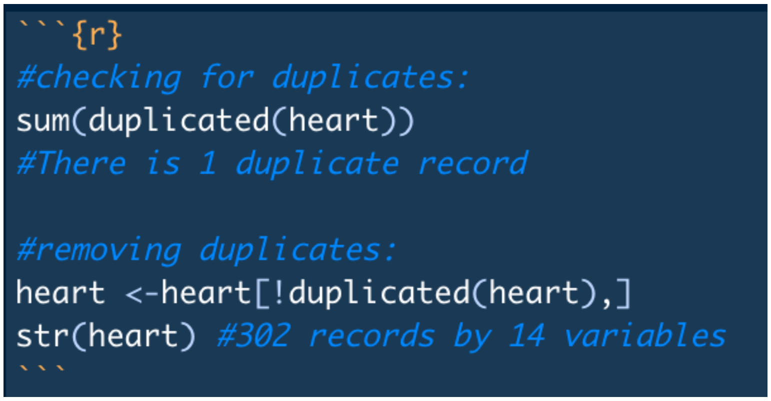
Beach[1]. The original data set contains 76 attributes, but a smaller version with 14 attributes was selected for this project. There are 303 observations. There are 14 variables in the dataset. ***Table 1*** below is the data dictionary with data types. ***Figure A.1*** in the Appendix shows a sample of the dataset in R. ***Figure A.2*** in the Appendix shows the summary statistics for each variable.

Attribute	**Description**	**Data Type**
Age	In years	Continuous
Sex	Gender: Male or Female	0 (F) and 1 (M)
Cp	Chest pain	0: typical angina 1: atypical angina 2: non-anginal pain 3: asymptomatic
Trestsbp	Resting blood pressure at hospital admission	Continuous
Chol	Serum cholesterol	Continuous
Fbs	Fasting blood sugar	0: Less than 120mg/dl fasting blood sugar 1: More than 120mg/dl fasting blood sugar
Restecg	Resting electrocardiographic results	0:normal 1:ST-T abnormality 2:Probable/Definite left ventricular hypertrophy
Thalach	Maximum heart rate achieved	Continuous
Exang	Exercise induced angina	0: no chest pain 1: chest pain
Oldpeak	ST depression caused by exercise	Continuous
Slope	Slope of ST depression	0: upsloping 1:flat 2:downsloping
Ca	Blood vessels colored by fluoroscopy	0-3

[1] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Thal	Disease in thallium test	0: absence 1: normal 2: fixed defect 3: reversible defect
Target	Presence of heart disease	0: high risk 1: low risk

Preprocessing:



~~~
```{r}
#checking for duplicates:
sum(duplicated(heart))
#There is 1 duplicate record

#removing duplicates:
heart <-heart[!duplicated(heart),]
str(heart) #302 records by 14 variables
```
~~~

All 14 variables were kept from the original dataset. There were no missing values and no NA values in the dataset. There was one duplicate record which was deleted. The dataset stands at 302 observations. Age, chol, and trestbps were discretized. Sex, cp, fbs, restecg, exang, slope, ca, thal and target were all converted from their respective data types to factors. The rest of the variables were re-coded. The overall quality of the data is more than satisfactory. All data types were structured and clean. None of the entries needed to be cleaned to remove symbols or unstructured text. ***Figures A.3*** and ***A.4*** in the Appendix show the data structure before and after cleaning. ***Figure 1*** to the left shows some of the preprocessing code.

Exploratory Data Analysis

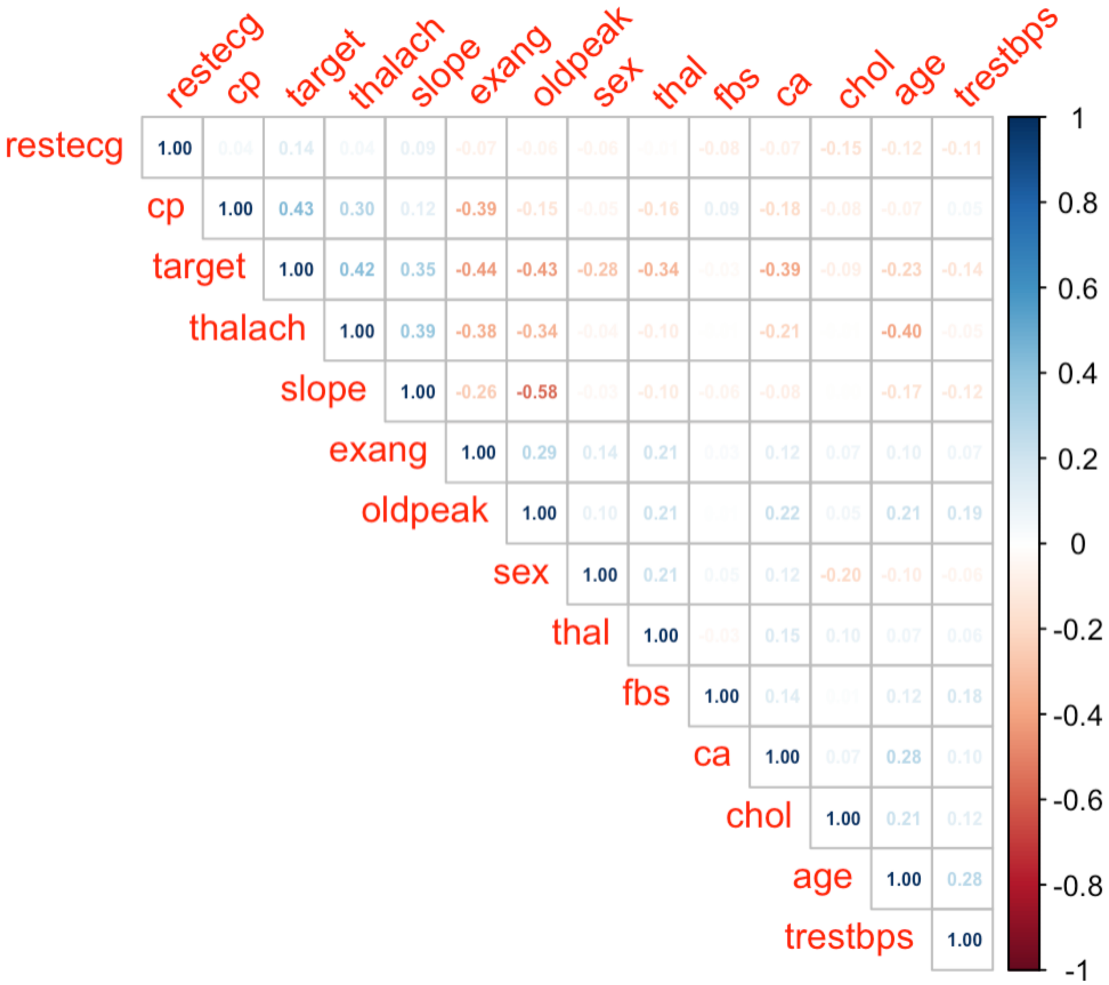
Figures A.5-18 in the Appendix show the distribution for all 14 attributes in the heart data set. The majority of the patients in the dataset are in their 40s and above. Only 6% of the patients are in their 20-30s. It follows the distribution of a bell curve, but is slightly left skewed. 68% of the patients are male. This is a bimodal distribution since there are only two factors. In terms of evidence for disease in the blood (thal), 55% have fixed defect, 38% have reservable defect, and 6% are normal. Thal is also left skewed with very few patients having normal or absence blood tests.

The majority of the patients either have typical angina chest pain (cp) or non-angina chest pain, with 143 and 86 patients respectively. Only 50 patients reported typical angina chest pain and 23 patients reported asymptomatic chest pain. Chest pain among the 302 patients is right skewed. 46% of patients reported prehypertension resting blood pressure (trestbp), 32% of patients had optimal resting blood pressure, 17% were in stage 1, 4% were in stage 2, and 0.6% were in hypertension crisis. Resting blood pressure is strongly right skewed with the majority of the patients belonging to either optimal or prehypertension resting blood pressure. 151 out of 302 patients were high risk in terms of their cholesterol levels. Cholesterol was left skewed. 85% of the patients had less than 120mg/dl in fasting blood sugar (fbs). Most of the patients were split between normal and ST-T wave abnormality (restecg) with only 4 out of 302 patients having probable/definite left ventricular hypertrophy.

The maximum heart rate achieved (thalach) among the patients was strongly left skewed. The majority had either high or normal with 8 out of 302 patients reporting lower than normal maximum heart rates. 67% of patients had no exercise induced angina. 68% had a slope of the peak for the ST depression greater than zero. The patients were split between flat and downsloping ST depression slopes, with only 7% having an upslope. The majority of patients have zero vessels colored by fluoroscopy at 58%.

Figures A.24-28 in the Appendix display boxplots of each continuous variable between low risk and high risk individuals, including age, restbps, chol, thalach, and oldpeak. The median age is roughly 52 for high risk patients, and 58 for low risk. The interquartile range for age is between roughly 44 and 59 for high risk individuals, while low risk is between roughly 52 and 62. The resting blood pressure (trestbps) shows the same median between low risk and high risk, at roughly 128, and the interquartile ranges are very similar, with low risk having a slightly wider range. The serum cholesterol levels (chol) show very similar medians and ranges between low risk and high risk individuals. There are more noticeable outliers showing high cholesterol on high risk individuals. The maximum heart rate achieved (thalach) reveals a much higher heart rate distribution for high risk individuals compared to low-risk. For oldpeak, the ST depression caused by exercise tends to be much lower for high-risk compared to low-risk patients.

Figure 2 shows the correlation matrix between all 14 variables.



	restecg	cp	target	thalach	slope	exang	oldpeak	sex	thal	fbs	ca	chol	age	trestbps
restecg	1.00	0.04	0.14	0.04	0.09	-0.07	-0.06	-0.06	-0.01	-0.08	-0.07	-0.15	-0.12	-0.11
cp		1.00	0.43	0.30	0.12	-0.39	-0.15	-0.05	-0.16	0.09	-0.18	-0.08	-0.07	0.05
target			1.00	0.42	0.35	-0.44	-0.43	-0.28	-0.34	-0.03	-0.39	-0.09	-0.23	-0.14
thalach				1.00	0.39	-0.38	-0.34	-0.04	-0.10	-0.01	-0.21	-0.01	-0.40	-0.05
slope					1.00	-0.26	-0.58	-0.03	-0.10	-0.06	-0.08	0.00	-0.17	-0.12
exang						1.00	0.29	0.14	0.21	0.03	0.12	0.07	0.10	0.07
oldpeak							1.00	0.10	0.21	0.01	0.22	0.05	0.21	0.19
sex								1.00	0.21	0.05	0.12	-0.20	-0.10	-0.06
thal									1.00	-0.03	0.15	0.10	0.07	0.06
fbs										1.00	0.14	0.01	0.12	0.18
ca											1.00	0.07	0.28	0.10
chol												1.00	0.21	0.12
age													1.00	0.28
trestbps														1.00

Three variables are strongly positively correlated with the target variable: cp (chest pain type) at r = 0.43, thalach (maximum heart rate achieved) at r = 0.42, and slope (slope at peak of ST) at r = 0.35. Five variables are negatively correlated: exang (exercise induced angina) at r = -0.44, oldpeak (ST depression induced by exercise) at r = -0.43, sex at r = -0.28, thal (inherited blood disorder) at r = -0.34, and Ca (# of major vessels colored by fluoroscopy) at r = -0.39.

Figures A.19-A.23 in the Appendix show barplots of the categorical variables with significant coefficients that distinguish between high and low risk. The five identifying variables are cp, exang, thal, slope, and thalach. One can presume that people with a high risk of heart disease would have Non-Angina (CP), no exercise induced angina (exang), fixed inherit blood disorder (thal), downsloping peak of ST exercise (slope) , and high maximum heart rate achieved (thalach).

Methods

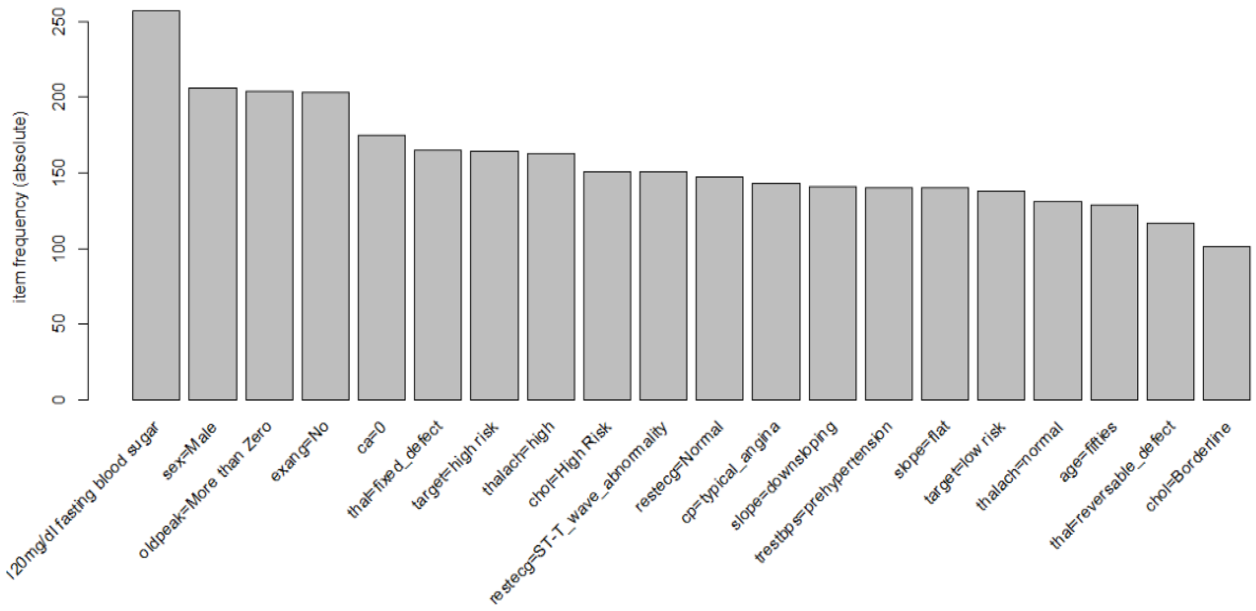
Association Rule Mining

Association Rule Mining (ARM) is an unsupervised machine learning technique that evaluates transactions to identify correlations and associations in categorical data. By applying ARM to the heart disease dataset, association rules were generated to determine what combination of factors are likely to result in a patient having heart disease. The following three measurement parameters are used in ARM:

- **Support:** The proportion of items occurring together relative to all transactions.
- **Confidence:** The proportion of items A and B occurring together relative to all transactions containing item A.
- **Lift:** The correlation between items A and B.

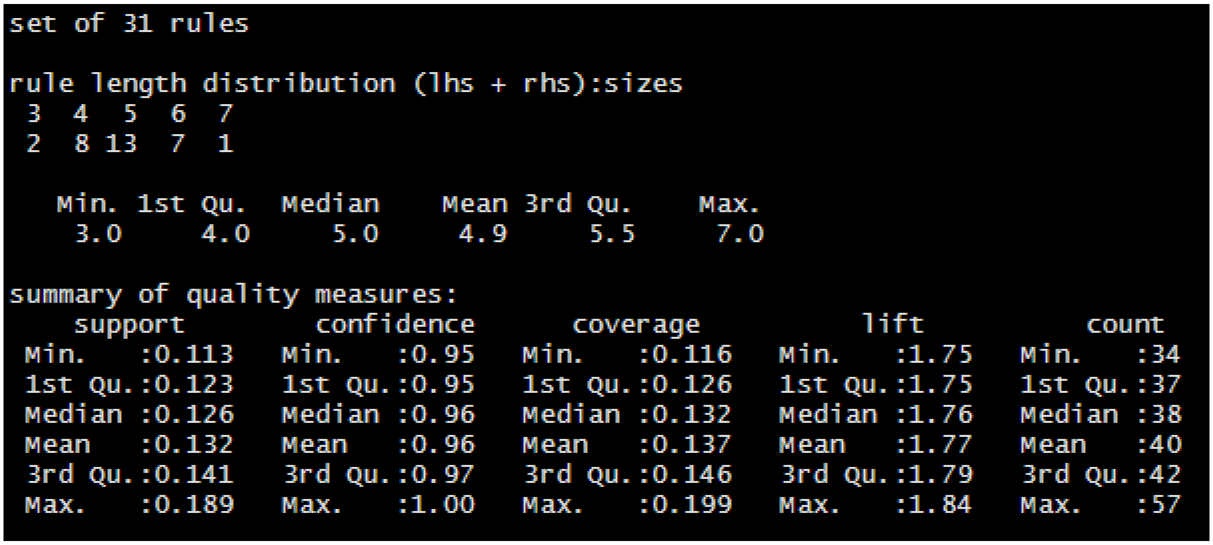
Association Rule Mining results should ideally contain strong rules with high lift, high confidence, and a relatively high support. In the heart disease dataset, the goal was to obtain between 20 and 40 strong rules. To achieve optimal results, minimal input measurements were set in the parameters, the model was run, results were evaluated, the input parameters were tweaked, and the model was rerun until the desired results were obtained.

Analysis: Before creating the model, the 20 most frequent items in the heart disease dataset were identified and displayed in the following item frequency plot (***Figure 3***).



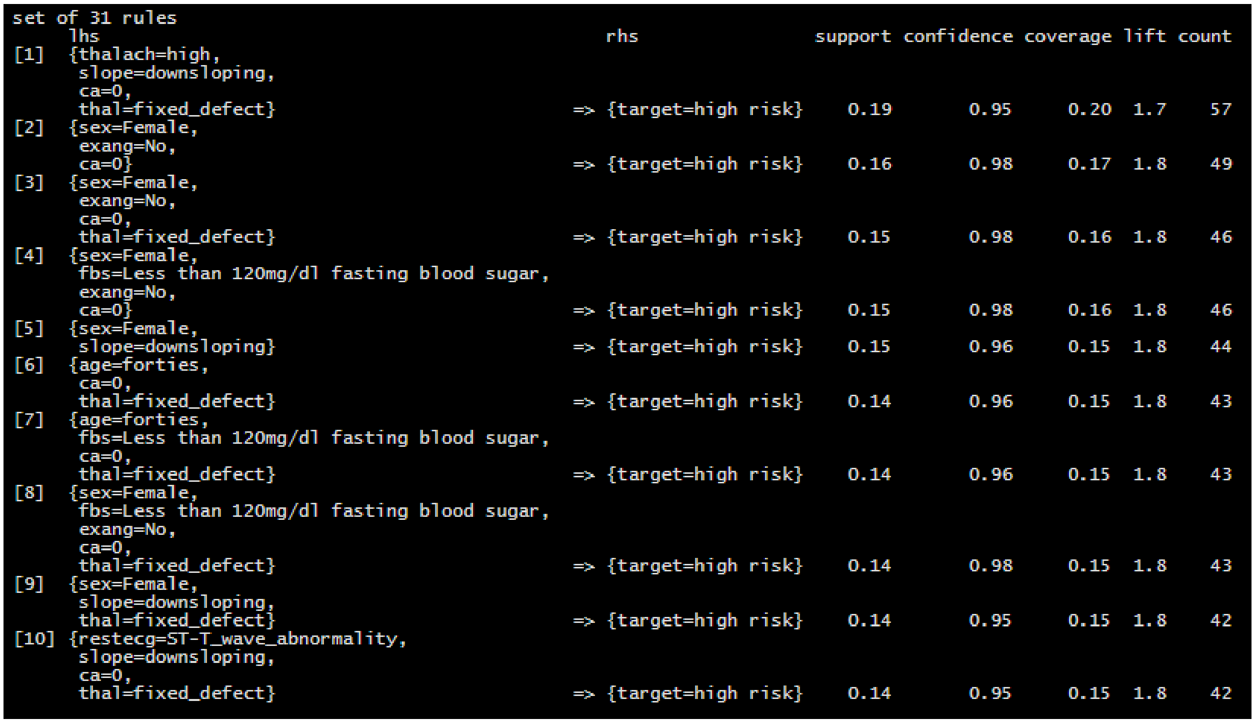
The apriori algorithm was used to create the ARM model. The right hand side was set to "target=high risk" to only generate rules that resulted in a patient being at high risk for heart disease. ***Figures B.1-2*** in the appendix describe and display the initial parameter inputs and results. After adjusting the parameters several times, the most optimal results were obtained with a minimum support of .11, a minimum confidence of .95, and a minimum length of 3 items. This resulted in 31 strong rules, as shown below.

Figure 4: **Rules Summary**



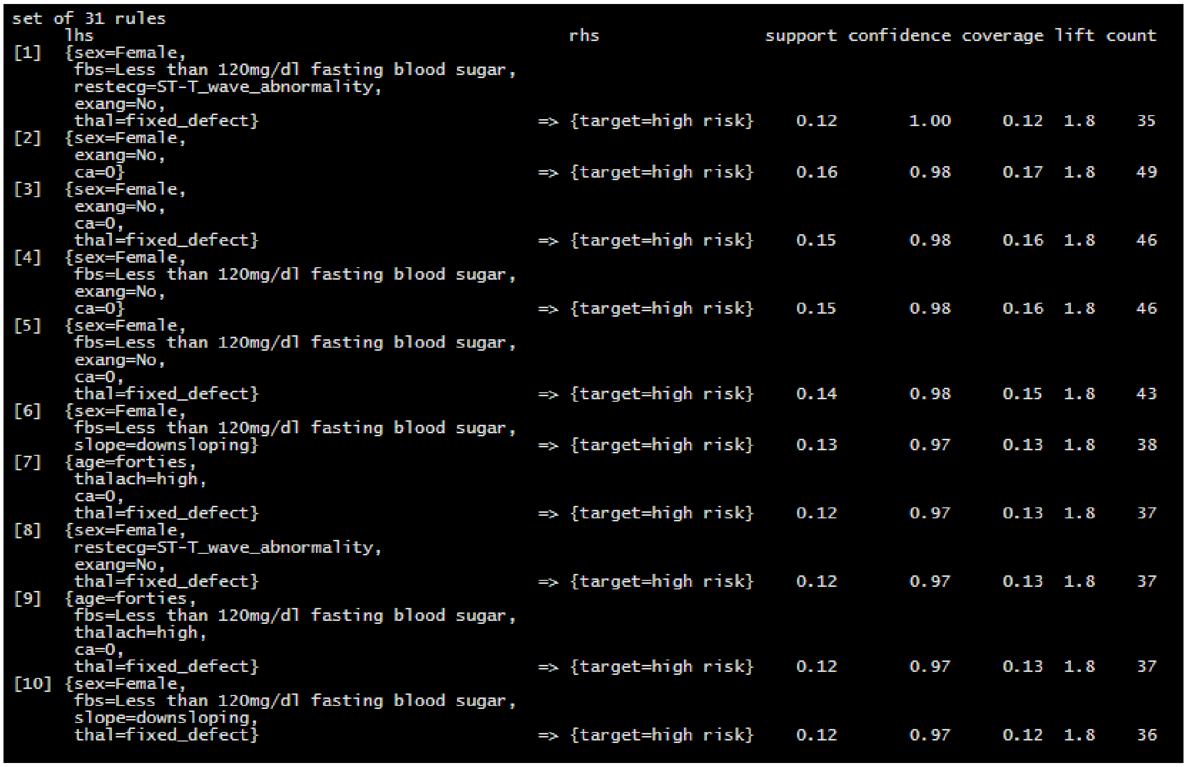
```
set of 31 rules

rule length distribution (lhs + rhs):sizes
 3  4  5  6  7 
 2  8 13  7  1 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    3.0     4.0     5.0     4.9     5.5     7.0 

summary of quality measures:
    support         confidence        coverage           lift           count   
 Min.   :0.113   Min.   :0.95   Min.   :0.116   Min.   :1.75   Min.   :34  
 1st Qu.:0.123   1st Qu.:0.95   1st Qu.:0.126   1st Qu.:1.75   1st Qu.:37  
 Median :0.126   Median :0.96   Median :0.132   Median :1.76   Median :38  
 Mean   :0.132   Mean   :0.96   Mean   :0.137   Mean   :1.77   Mean   :40  
 3rd Qu.:0.141   3rd Qu.:0.97   3rd Qu.:0.146   3rd Qu.:1.79   3rd Qu.:42  
 Max.   :0.189   Max.   :1.00   Max.   :0.199   Max.   :1.84   Max.   :57  
```

Figure 5: **Top 10 Strong Rules, Ordered By Support Descending**



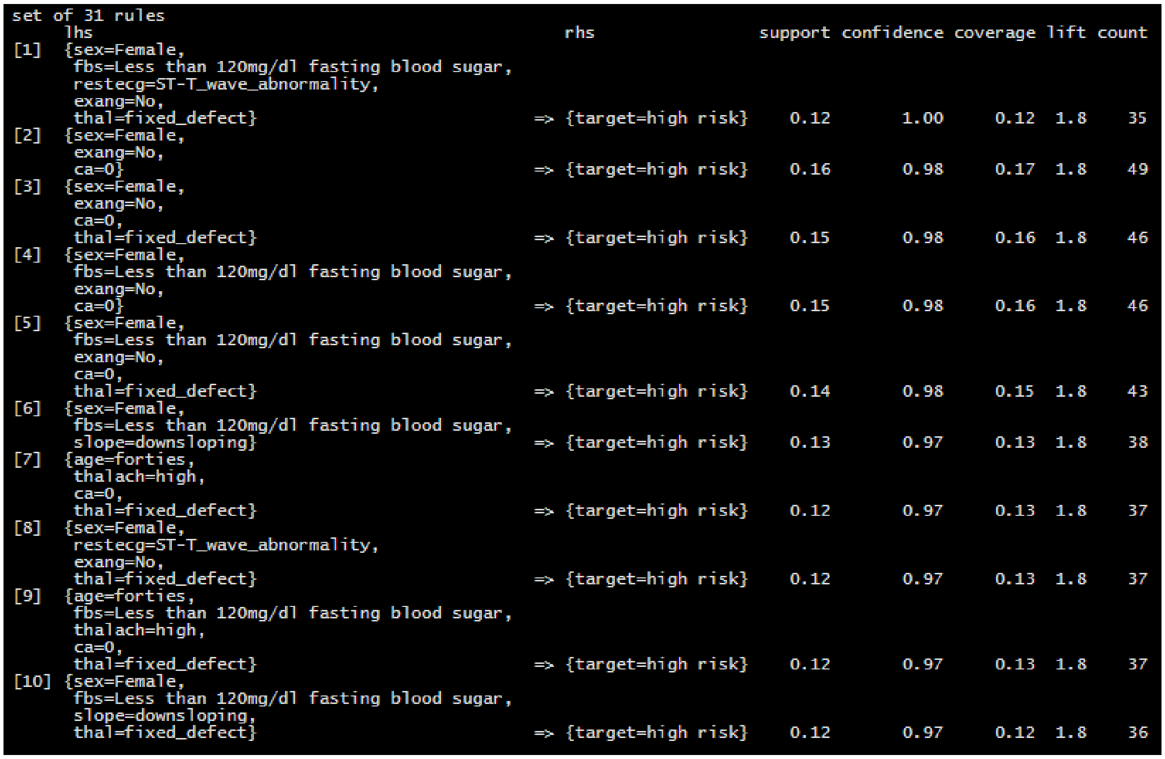
```
set of 31 rules
     lhs                                               rhs                 support confidence coverage lift count
[1]  {thalach=high,
      slope=downsloping,
      ca=0,
      thal=fixed_defect}                            => {target=high risk}    0.19       0.95     0.20  1.7    57
[2]  {sex=Female,
      exang=No,
      ca=0}                                         => {target=high risk}    0.16       0.98     0.17  1.8    49
[3]  {sex=Female,
      exang=No,
      ca=0,
      thal=fixed_defect}                            => {target=high risk}    0.15       0.98     0.16  1.8    46
[4]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      exang=No,
      ca=0}                                         => {target=high risk}    0.15       0.98     0.16  1.8    46
[5]  {sex=Female,
      slope=downsloping}                            => {target=high risk}    0.15       0.96     0.15  1.8    44
[6]  {age=forties,
      ca=0,
      thal=fixed_defect}                            => {target=high risk}    0.14       0.96     0.15  1.8    43
[7]  {age=forties,
      fbs=Less than 120mg/dl fasting blood sugar,
      ca=0,
      thal=fixed_defect}                            => {target=high risk}    0.14       0.96     0.15  1.8    43
[8]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      exang=No,
      ca=0,
      thal=fixed_defect}                            => {target=high risk}    0.14       0.98     0.15  1.8    43
[9]  {sex=Female,
      slope=downsloping,
      thal=fixed_defect}                            => {target=high risk}    0.14       0.95     0.15  1.8    42
[10] {restecg=ST-T_wave_abnormality,
      slope=downsloping,
      ca=0,
      thal=fixed_defect}                            => {target=high risk}    0.14       0.95     0.15  1.8    42
```

Figure 6: **Top 10 Strong Rules, Ordered By Confidence Descending**



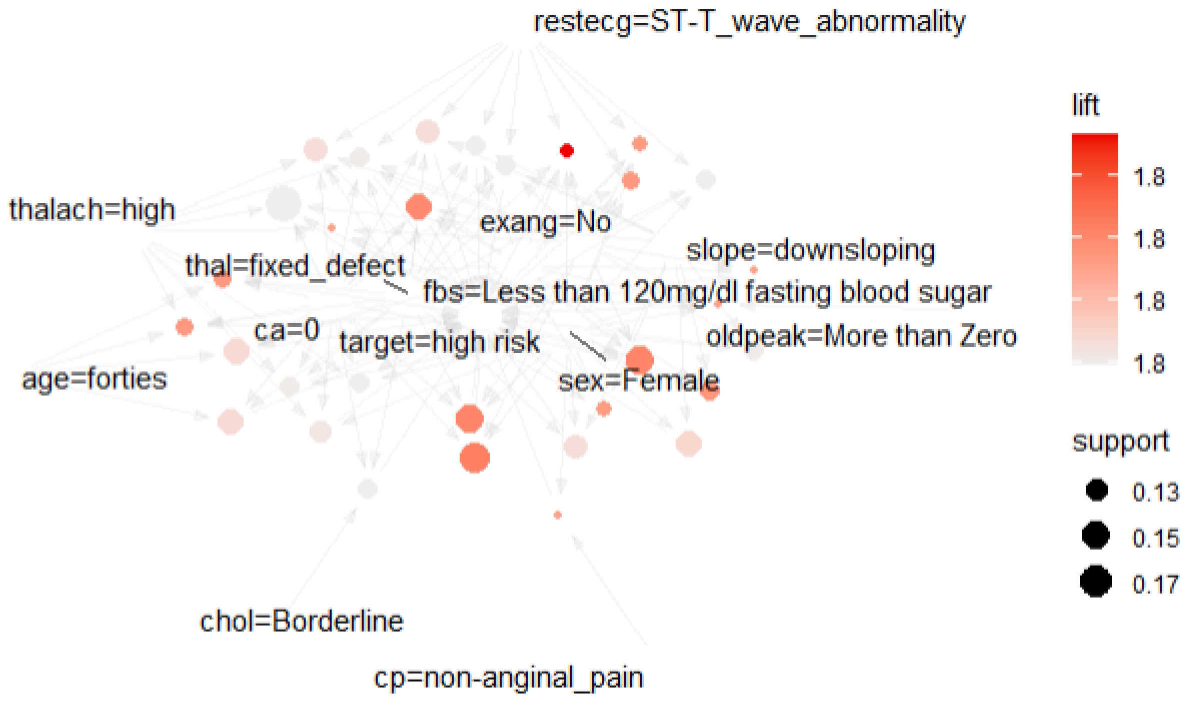
```
set of 31 rules
     lhs                                                rhs               support confidence coverage lift count
[1]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      restecg=ST-T_wave_abnormality,
      exang=No,
      thal=fixed_defect}                             => {target=high risk}    0.12       1.00     0.12  1.8    35
[2]  {sex=Female,
      exang=No,
      ca=0}                                          => {target=high risk}    0.16       0.98     0.17  1.8    49
[3]  {sex=Female,
      exang=No,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.15       0.98     0.16  1.8    46
[4]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      exang=No,
      ca=0}                                          => {target=high risk}    0.15       0.98     0.16  1.8    46
[5]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      exang=No,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.14       0.98     0.15  1.8    43
[6]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      slope=downsloping}                             => {target=high risk}    0.13       0.97     0.13  1.8    38
[7]  {age=forties,
      thalach=high,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.13  1.8    37
[8]  {sex=Female,
      restecg=ST-T_wave_abnormality,
      exang=No,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.13  1.8    37
[9]  {age=forties,
      fbs=Less than 120mg/dl fasting blood sugar,
      thalach=high,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.13  1.8    37
[10] {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      slope=downsloping,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.12  1.8    36
```

Figure 7: **Top 10 Strong Rules, Ordered By Lift Descending**



```
set of 31 rules
     lhs                                                rhs               support confidence coverage lift count
[1]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      restecg=ST-T_wave_abnormality,
      exang=No,
      thal=fixed_defect}                             => {target=high risk}    0.12       1.00     0.12  1.8    35
[2]  {sex=Female,
      exang=No,
      ca=0}                                          => {target=high risk}    0.16       0.98     0.17  1.8    49
[3]  {sex=Female,
      exang=No,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.15       0.98     0.16  1.8    46
[4]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      exang=No,
      ca=0}                                          => {target=high risk}    0.15       0.98     0.16  1.8    46
[5]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      exang=No,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.14       0.98     0.15  1.8    43
[6]  {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      slope=downsloping}                             => {target=high risk}    0.13       0.97     0.13  1.8    38
[7]  {age=forties,
      thalach=high,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.13  1.8    37
[8]  {sex=Female,
      restecg=ST-T_wave_abnormality,
      exang=No,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.13  1.8    37
[9]  {age=forties,
      fbs=Less than 120mg/dl fasting blood sugar,
      thalach=high,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.13  1.8    37
[10] {sex=Female,
      fbs=Less than 120mg/dl fasting blood sugar,
      slope=downsloping,
      thal=fixed_defect}                             => {target=high risk}    0.12       0.97     0.12  1.8    36
```

Figure 8: **31 Strong Rules Plot**



Interesting Rules

Of the 31 strong rules, five interesting rules have been selected, and can be summarized as follows:

Table 2: **Rule 1:**

LHS	RHS	Support	Confidence	Coverage	Lift	Count
chol=Borderline, ca=0, thal=fixed_defect	Target = High Risk	.13	.95	.013	1.7	38

- **Interpretation:** A patient who has borderline between low-risk and high-risk cholesterol, zero blood vessels colored by fluoroscopy, and a thallium test resulting in fixed-defect is at risk for heart disease.

Table 3: **Rule 2:**

LHS	RHS	Support	Confidence	Coverage	Lift	Count
age=forties, thalach=high, ca=0, thal=fixed_defect	Target = High Risk	.12	.97	.013	1.8	37

- **Interpretation:** A patient in their forties with zero blood vessels colored by fluoroscopy, a thallium test resulting in fixed-defect, and a maximum heart rate in the high range is at risk for heart disease.

Table 4: **Rule 3:**

LHS	RHS	Support	Confidence	Coverage	Lift	Count
sex=Female, thalach=high, ca=0	Target = High Risk	.13	.95	.014	1.8	40

- **Interpretation:** A female patient who has a maximum heart rate in the high range and zero blood vessels colored by fluoroscopy is at risk for heart disease.

Table 5: **Rule 4:**

LHS	RHS	Support	Confidence	Coverage	Lift	Count
sex=Female, slope=downsloping	Target = High Risk	.15	.96	.015	1.8	44

- **Interpretation:** A female patient who has a downsloping ST depression is at risk for heart disease.

Table 6: **Rule 5:**

LHS	RHS	Support	Confidence	Coverage	Lift	Count
restecg=ST-T_wave_abnormality, slope=downsloping, ca=0, thal=fixed_defect	Target = High Risk	.14	.95	.015	1.8	42

- **Interpretation:** A patient with a resting electrocardiographic result of ST-T abnormality, a downward sloping ST depression, a thallium test resulting in fixed-defect, and zero blood vessels colored by fluoroscopy is at risk for heart disease.

Decision Tree

Decision trees are used to evaluate possible outcomes and their chances. Within a decision tree, the nodes represent a single variable. A leaf represents a move to another node. The initial node is called the root.

Figure 7: **Unpruned decision tree**

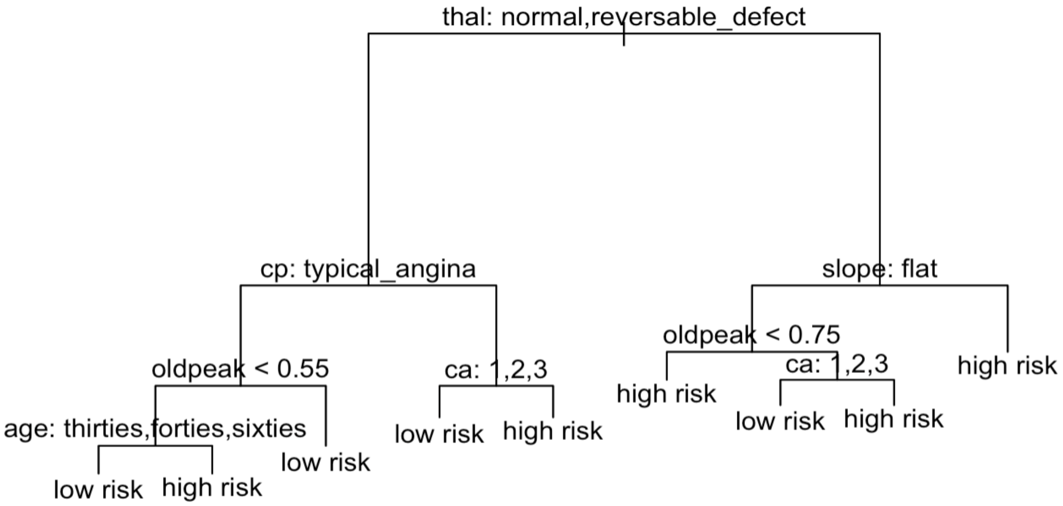
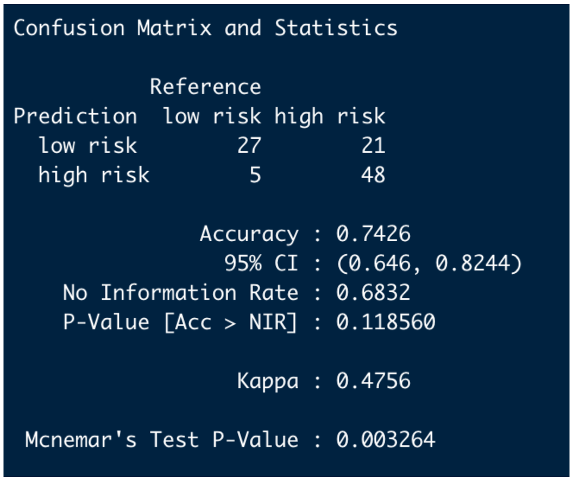


Figure 7.A: **Confusion matrix for unpruned decision tree**



```
Confusion Matrix and Statistics

           Reference
Prediction  low risk high risk
  low risk        27        21
  high risk        5        48

               Accuracy : 0.7426
                 95% CI : (0.646, 0.8244)
    No Information Rate : 0.6832
    P-Value [Acc > NIR] : 0.118560

                  Kappa : 0.4756

 Mcnemar's Test P-Value : 0.003264
```

Analysis: Figure 7 above shows the initial node and the general tree isn't strong enough, with an accuracy of 74.26% shown in the confusion matrix in ***Figure 7.A.***

Figure 8: **Pruned Decision Tree**

Therefore, a pruned decision tree was created and shown in ***Figure 8***. The pruned decision tree has an accuracy of 81.19%. This is a great improvement from the initial decision tree.

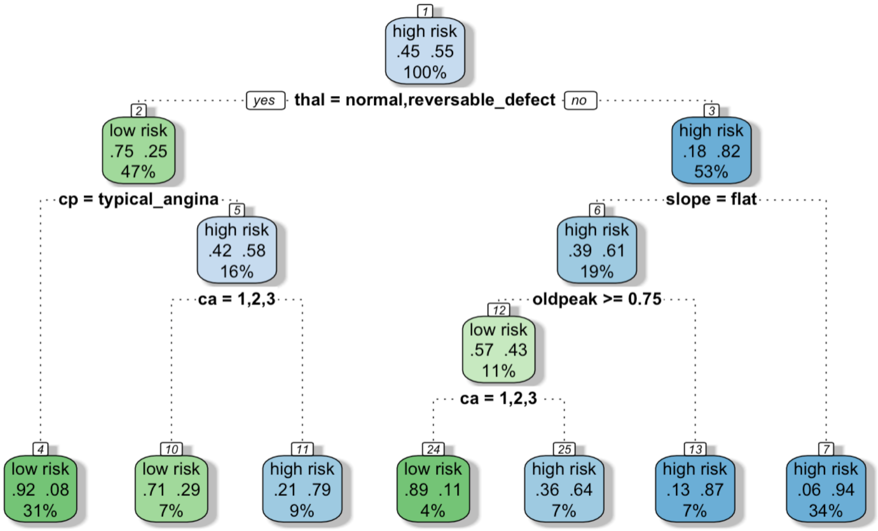
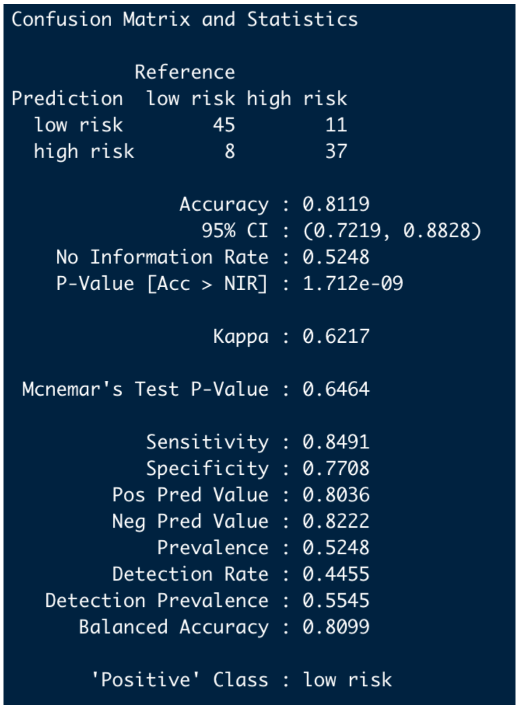


Figure 8.A: **Confusion Matrix for Pruned Decision Tree**



```
Confusion Matrix and Statistics

           Reference
Prediction  low risk high risk
  low risk        45        11
  high risk        8        37

               Accuracy : 0.8119
                 95% CI : (0.7219, 0.8828)
    No Information Rate : 0.5248
    P-Value [Acc > NIR] : 1.712e-09

                  Kappa : 0.6217

 Mcnemar's Test P-Value : 0.6464

            Sensitivity : 0.8491
            Specificity : 0.7708
         Pos Pred Value : 0.8036
         Neg Pred Value : 0.8222
             Prevalence : 0.5248
         Detection Rate : 0.4455
   Detection Prevalence : 0.5545
      Balanced Accuracy : 0.8099

       'Positive' Class : low risk
```

Figure 9: **Decision tree with high correlation variables**

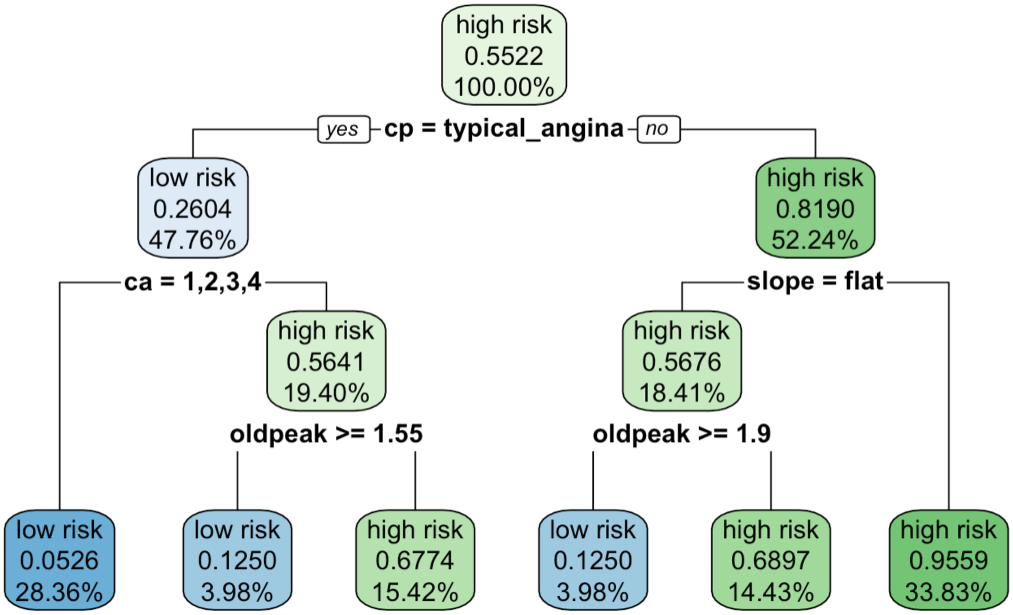
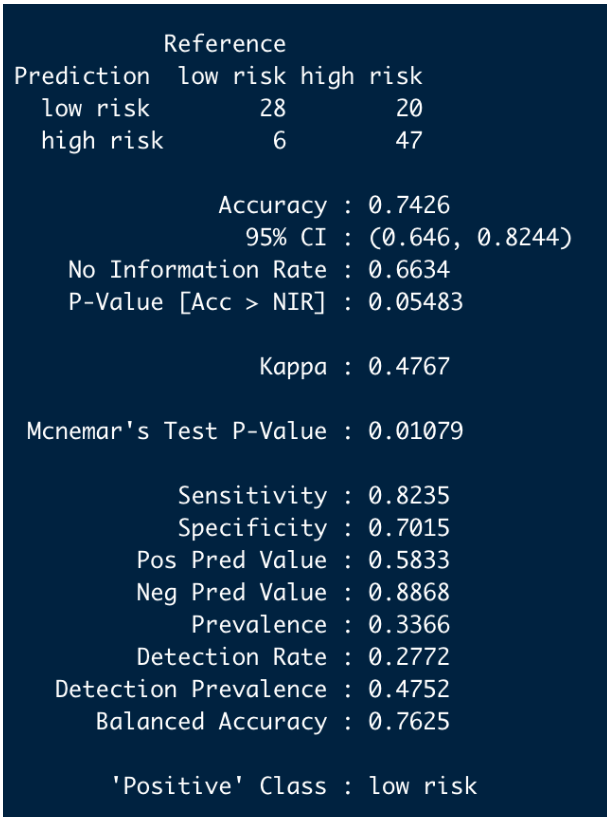


Figure 9.A: **Confusion Matrix for Decision Tree with high correlation variables**



```
           Reference
Prediction  low risk high risk
  low risk        28        20
  high risk        6        47
                                          
               Accuracy : 0.7426          
                 95% CI : (0.646, 0.8244)
    No Information Rate : 0.6634          
    P-Value [Acc > NIR] : 0.05483         
                                          
                  Kappa : 0.4767          
                                          
 Mcnemar's Test P-Value : 0.01079         
                                          
            Sensitivity : 0.8235          
            Specificity : 0.7015          
         Pos Pred Value : 0.5833          
         Neg Pred Value : 0.8868          
             Prevalence : 0.3366          
         Detection Rate : 0.2772          
   Detection Prevalence : 0.4752          
      Balanced Accuracy : 0.7625          
                                          
       'Positive' Class : low risk        
```

Figure 9 above shows the decision tree model made for the variables with high correlation. The accuracy of this tree was 74.26% as shown below in the confusion matrix (***Figure 9.A***)

Figure 10: **Decision tree for sex, age, cholestrol and blood pressure**

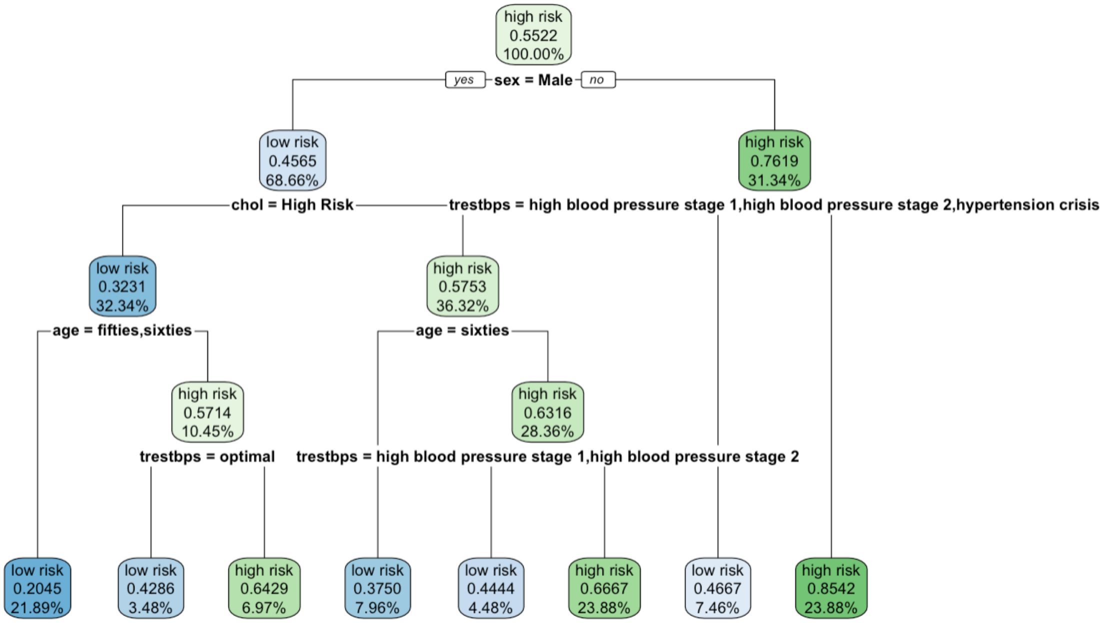
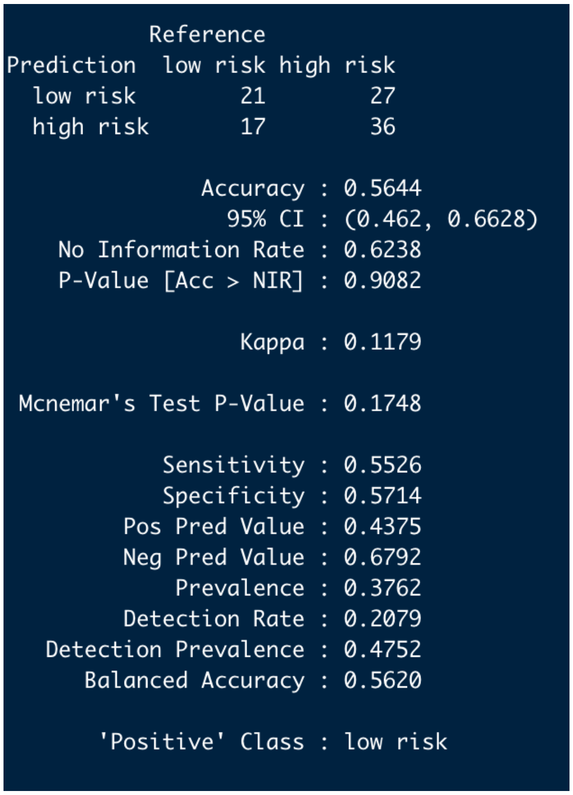


***Figure 10.A:* Confusion matrix**



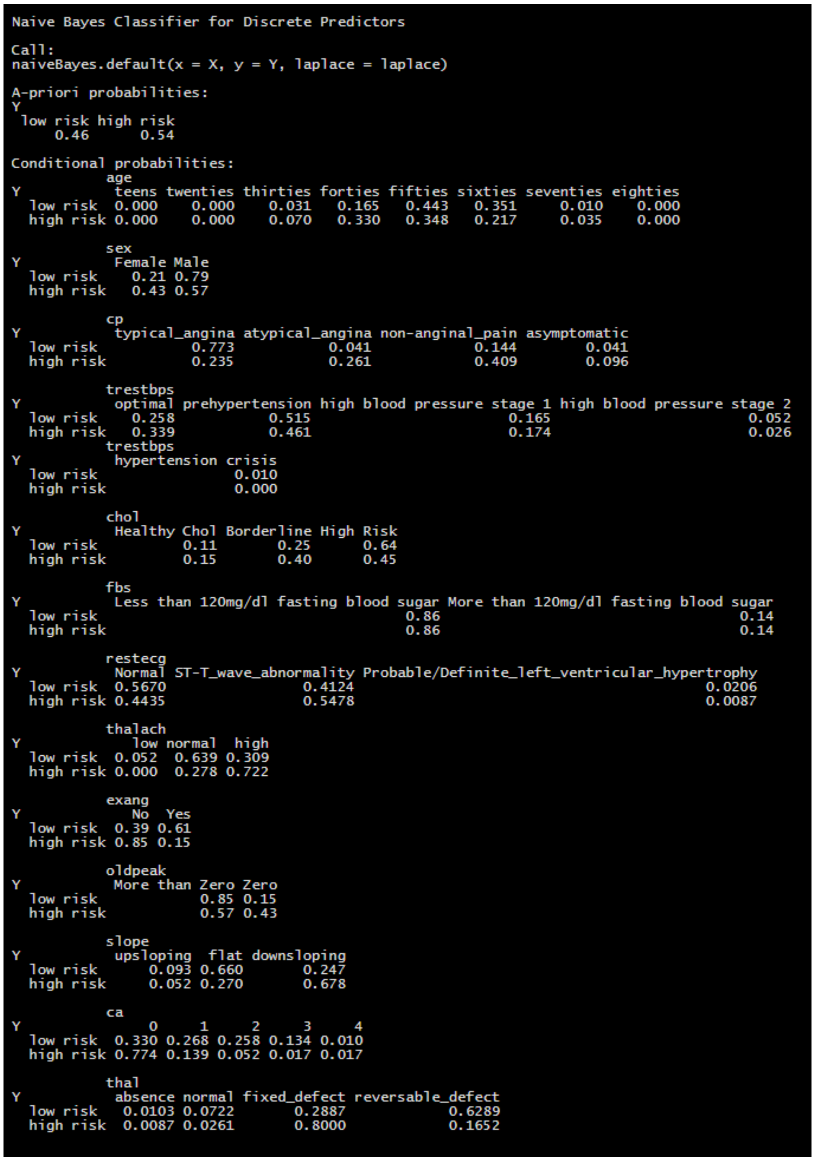
```
           Reference
Prediction  low risk high risk
  low risk        21        27
  high risk       17        36

               Accuracy : 0.5644
                 95% CI : (0.462, 0.6628)
    No Information Rate : 0.6238
    P-Value [Acc > NIR] : 0.9082

                  Kappa : 0.1179

 Mcnemar's Test P-Value : 0.1748

            Sensitivity : 0.5526
            Specificity : 0.5714
         Pos Pred Value : 0.4375
         Neg Pred Value : 0.6792
             Prevalence : 0.3762
         Detection Rate : 0.2079
   Detection Prevalence : 0.4752
      Balanced Accuracy : 0.5620

       'Positive' Class : low risk
```

Figure 10 above shows the decision tree model made for the variables for demographics such as age and sex, and heart measures such as cholestrol and blood pressure. The accuracy for this tree is 56.44%, shown in Figure ***10.A***.

Naïve Bayes

The Naïve Bayes Classifier (NBC) is a classification algorithm that predicts distributions over a set of outcomes based on probability. In the case of the heart disease data, each patient was classified as either being at risk for heart disease or not based on conditional probabilities between the target variable and the predictor variables. This classification model assumes that there is conditional independence among the variables. Based on the previously illustrated correlation matrix, there were not overly strong correlations between the variables. This adds confirmation to the validity of this model.

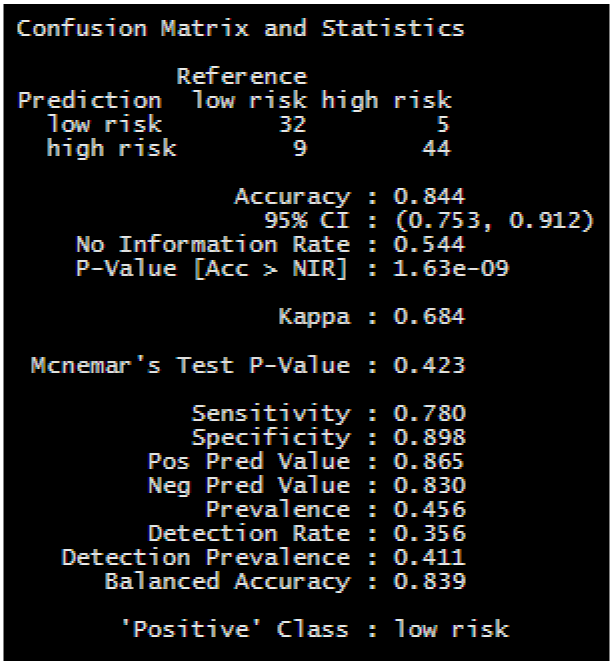
Analysis: The following visual displays the results from the Naïve Bayes training model, including the conditional probabilities between the target variable and the independent variables seen in ***Figure 11*** below.



```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
 low risk high risk 
     0.46      0.54 

Conditional probabilities:
           age
Y           teens twenties thirties forties fifties sixties seventies eighties
  low risk  0.000    0.000    0.031   0.165   0.443   0.351     0.010    0.000
  high risk 0.000    0.000    0.070   0.330   0.348   0.217     0.035    0.000

           sex
Y           Female Male
  low risk    0.21 0.79
  high risk   0.43 0.57

           cp
Y           typical_angina atypical_angina non-anginal_pain asymptomatic
  low risk           0.773           0.041            0.144        0.041
  high risk          0.235           0.261            0.409        0.096

           trestbps
Y           optimal prehypertension high blood pressure stage 1 high blood pressure stage 2
  low risk    0.258           0.515                       0.165                       0.052
  high risk   0.339           0.461                       0.174                       0.026
           trestbps
Y           hypertension crisis
  low risk                0.010
  high risk               0.000

           chol
Y           Healthy Chol Borderline High Risk
  low risk          0.11       0.25      0.64
  high risk         0.15       0.40      0.45

           fbs
Y           Less than 120mg/dl fasting blood sugar More than 120mg/dl fasting blood sugar
  low risk                                    0.86                                   0.14
  high risk                                   0.86                                   0.14

           restecg
Y           Normal ST-T_wave_abnormality Probable/Definite_left_ventricular_hypertrophy
  low risk  0.5670                0.4124                                         0.0206
  high risk 0.4435                0.5478                                         0.0087

           thalach
Y              low normal  high
  low risk  0.052  0.639 0.309
  high risk 0.000  0.278 0.722

           exang
Y              No  Yes
  low risk  0.39 0.61
  high risk 0.85 0.15

           oldpeak
Y           More than Zero Zero
  low risk            0.85 0.15
  high risk           0.57 0.43

           slope
Y           upsloping  flat downsloping
  low risk      0.093 0.660       0.247
  high risk     0.052 0.270       0.678

           ca
Y                0     1     2     3     4
  low risk  0.330 0.268 0.258 0.134 0.010
  high risk 0.774 0.139 0.052 0.017 0.017

           thal
Y           absence normal fixed_defect reversable_defect
  low risk   0.0103 0.0722       0.2887            0.6289
  high risk  0.0087 0.0261       0.8000            0.1652
```

After running the Naïve Bayes model against the test dataset, the model correctly predicted a patient being at high risk or low risk for heart disease with 84.4% accuracy. It correctly predicted 32 low risk patients, and 44 high risk patients. There were 9 false positives, and 5 false negatives. This is illustrated in the confusion matrix below.

Figure 12: **Naïve Bayes Confusion Matrix**



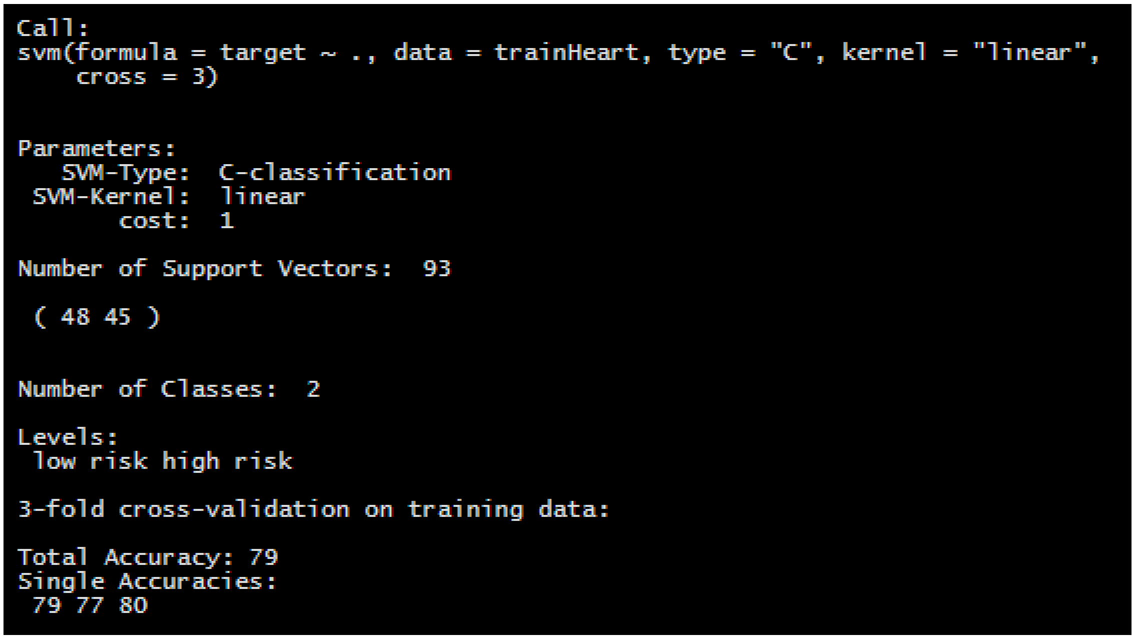
```
Confusion Matrix and Statistics

           Reference
Prediction  low risk high risk
  low risk        32         5
  high risk        9        44

               Accuracy : 0.844
                 95% CI : (0.753, 0.912)
    No Information Rate : 0.544
    P-Value [Acc > NIR] : 1.63e-09

                  Kappa : 0.684

 Mcnemar's Test P-Value : 0.423

            Sensitivity : 0.780
            Specificity : 0.898
         Pos Pred Value : 0.865
         Neg Pred Value : 0.830
             Prevalence : 0.456
         Detection Rate : 0.356
   Detection Prevalence : 0.411
      Balanced Accuracy : 0.839

       'Positive' Class : low risk
```

Support Vector Machine

Support Vector Machines (SVMs) are supervised machine learning models. These models can create both linear and nonlinear classifications. The goal is to create a maximum gap between the classification categories.

Analysis: The SVM model created on the heart disease dataset was trained with 3-fold cross-validation using a linear kernel. ***Figure 13*** below shows nearly an 80% total accuracy on the 3 folds.

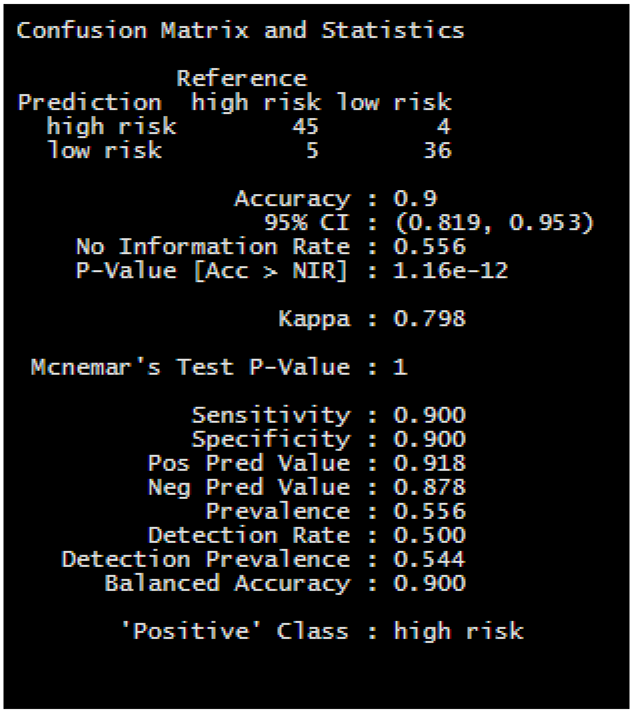
Figure 13: **Support Vector Machine 3-Fold Cross-Validation**



```
Call:
svm(formula = target ~ ., data = trainHeart, type = "C", kernel = "linear",
    cross = 3)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  93

 ( 48 45 )


Number of Classes:  2

Levels:
 low risk high risk

3-fold cross-validation on training data:

Total Accuracy: 79
Single Accuracies:
 79 77 80
```

When running the prediction model on the test set, the results came out with a 90% accuracy. There were 4 false positives and 5 false negatives, as shown in ***Figure 14*** below..

Figure 14: **SVM Confusion Matrix**



```
Confusion Matrix and Statistics

           Reference
Prediction  high risk low risk
  high risk        45        4
  low risk          5       36

               Accuracy : 0.9
                 95% CI : (0.819, 0.953)
    No Information Rate : 0.556
    P-Value [Acc > NIR] : 1.16e-12

                  Kappa : 0.798

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.900
            Specificity : 0.900
         Pos Pred Value : 0.918
         Neg Pred Value : 0.878
             Prevalence : 0.556
         Detection Rate : 0.500
   Detection Prevalence : 0.544
      Balanced Accuracy : 0.900

       'Positive' Class : high risk
```

Results

Association Rule Mining

Based on the strong rules that were mined from association rule mining ***(Table 2-6)***, some of the common variables associated with high risk heart disease patients include the following:

- Sex of female
- Fluoroscopy revealing no colored blood vessels
- Downsloping ST depression
- Thallium test resulting in fixed defect
- Resting electrocardiographic results revealing ST-T abnormality
- In the forties age group

When a patient is examined for heart disease, these factors can be explored as contributing factors based on the associated rules mined.

Decision Tree

Decision tree is a supervised machine learning method. The results from the unpruned decision tree showed that patients have a higher heart disease risk if they did not have normal or reversible defect in their thallium test (thalach) and did have flat slope for their slope of ST depression (slope) seen in ***Figure 8***. The highest accuracy is present in the pruned decision tree at 84.91%. The decision tree with the demographic variables showed that male patients had the highest risk of heart disease seen in ***Figure 10.*** The decision tree containing only the highest correlation variables showed that the patients with typical angina and a flat ST depression slope as having high risk of heart disease seen in ***Figure 9.***

Naïve Bayes

Naïve Bayes is also a supervised machine learning method. Some noteable conditional probabilities of note that were calculated from the model are as follows:

- The probability of a patient being at high risk for heart disease, given the gender of male is 57%. Conversely, the probability of a patient being high risk, given the gender of female, is 43%.
- The probability of a patient being at high risk for heart disease, given cholesterol falls in the high risk range, is 45%.
- The probability of a patient being at high risk for heart disease, given they have zero blood vessels colored by fluoroscopy, is 77%.
- The probability of a patient being at high risk for heart disease, given a thallium test resulting in fixed-defect, is 80%.

Support Vector Machine

As seen in ***Figure 14***, The Support Vector Machine model had a very high accuracy of 90%. This was the highest accuracy out of all of the models that were used. Despite the high accuracy, it could not determine the importance of the different variables in predicting the outcome of high risk or low risk patients.

Conclusion

Table 7:

Model	**Accuracy**	**Significant Variables**
ARM	Not applicable	Sex, age, major vessels colored by fluoroscoopy, thallium test, electrocardiographic results, and slope of ST depression,
Decision Tree	81.19%	Thallium test and slope of ST depression
Naïve Bayes	84.4%	Thallium test, major vessels colored by fluoroscopy, and sex
SVM	90%	Not Applicable

In conclusion, this research can provide interesting insights to those who are concerned about heart disease. The most commonly known symptom of heart disease is chest pain, but the machine learning models have shown another facet of heart disease. The results from association rule mining (ARM), the only unsupervised technique used, showed several interesting rules, such as the significant variables listed in ***Table 7.*** The supervised machine learning methods also identified significant variables, as well as the ability to observe the accuracy of the predictions. In terms of accuracy, the most successful data mining technique was the Support Vector Machine (SVM) model, with a 90% accuracy, followed closely by the decision tree and Naïve Bayes models, with 81.19% and 84.4% accuracies, respectively. The one common significant variable identified between the Naïve Bayes, decision tree, and ARM techniques was the thallium test, which measures how much blood is reaching different parts of the heart. Additionally, the Naïve Bayes and ARM models had major blood vessels colored by fluoroscopy and sex in common as significant variables. Between the decision tree and ARM models, the slope of ST depression was another common significant variable for predicting heart disease.

These results can be used as an educational source for doctors and patients who are interested in understanding heart disease predictions. While these models used on the heart disease data set closely predicted the presence of heart disease, the same algorithms can be a powerful tool for medical professionals in other fields of medicine to better research and understand various diseases. Machine learning methods could also be used on smaller data sets for patients with more rare diseases. While rare diseases are harder to research, machine learning techniques used in this project could provide researchers with a strong starting point.

The future for both supervised and unsupervised machine learning is boundless. This project shows the power of these techniques to take a data set and output interesting insights within medicine. In addition to medicine, machine learning can be applied to any industry. For example, machine learning has the power to process unstructured data, and can even distinguish images of dogs from images of cats. As of March 2022, researchers at Texas A&M AgriLife Research used machine learning to identify algae as a potential new source of biofuel. These, among many more examples, demonstrate the value of machine learning to help improve human life.

References

Centers for Disease Control and Prevention. (2022, January 13). *FASTSTATS - leading causes of death*. Centers for Disease Control and Prevention. Retrieved March 9, 2022, from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Centers for Disease Control and Prevention. (2022, February 7). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved March 9, 2022, from https://www.cdc.gov/heartdisease/facts.htm

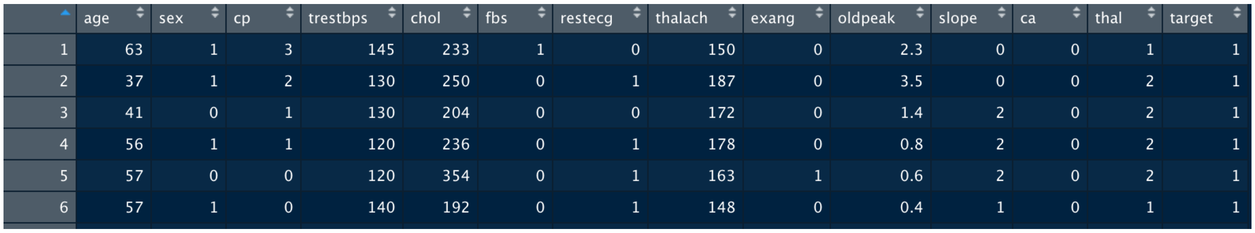
Types of heart disease. Heart and Stroke Foundation of Canada. (n.d.). Retrieved March 9, 2022, from https://www.heartandstroke.ca/heart-disease/what-is-heart-disease/types-of-heart-disease

World Health Organization. (n.d.). *The top 10 causes of death*. World Health Organization. Retrieved March 9, 2022, from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

World Health Organization. (n.d.). *Cardiovascular diseases (cvds)*. World Health Organization. Retrieved March 9, 2022, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

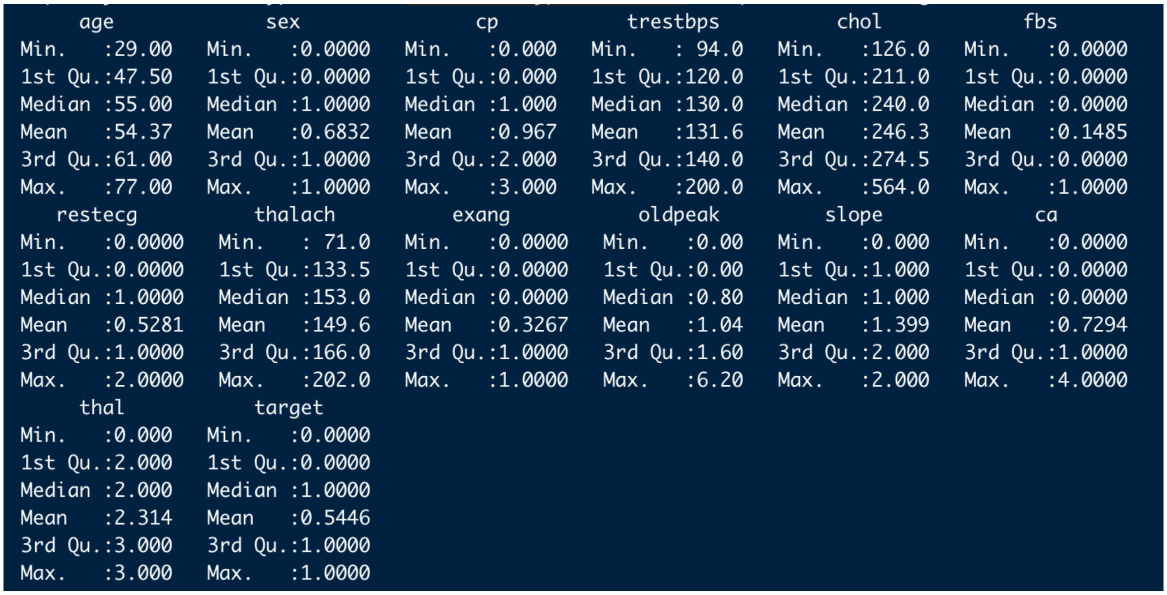
Appendix

Figure A.1



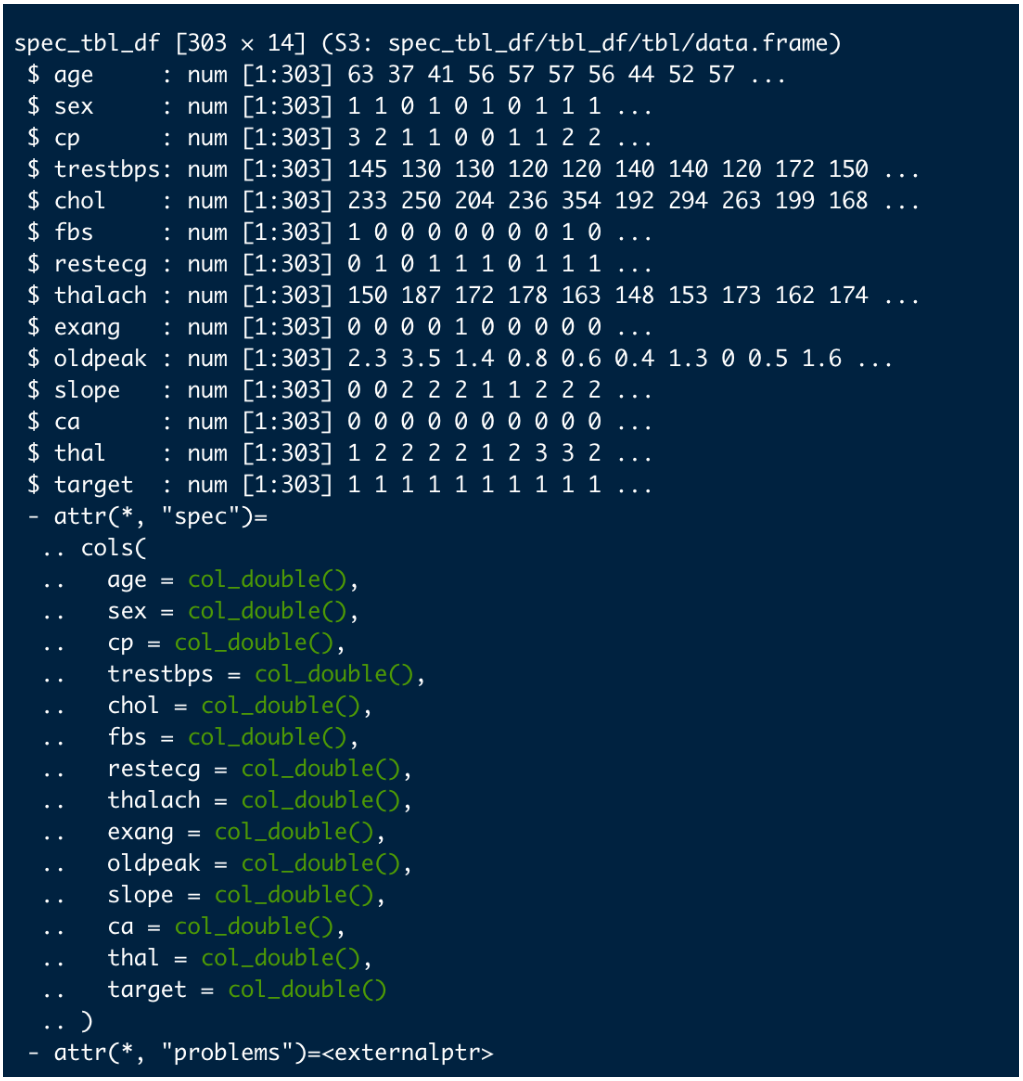
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

Figure A.2



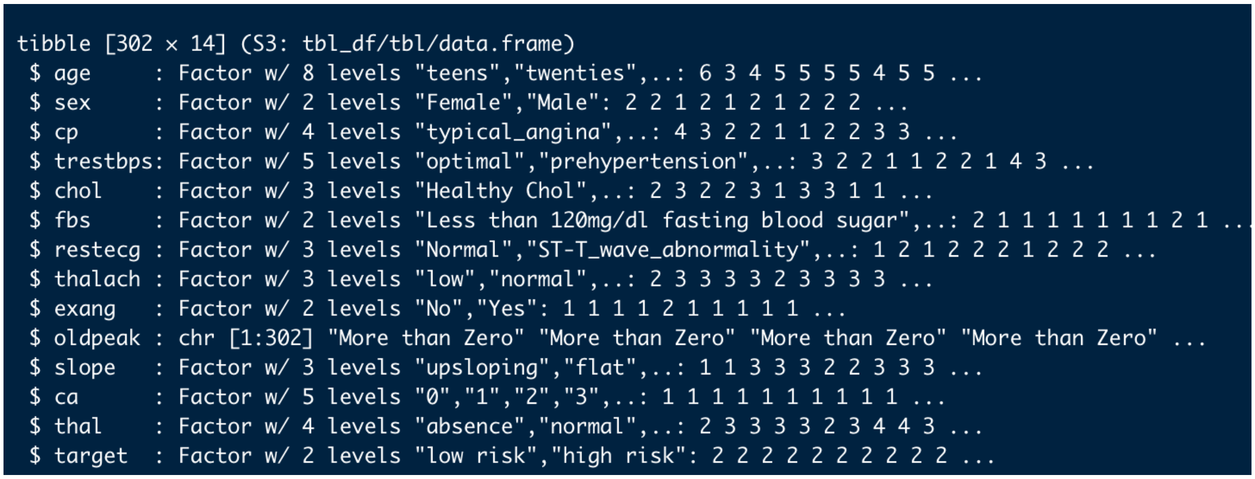
```
      age             sex               cp            trestbps          chol             fbs        
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000  
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000  
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0   Median :0.0000  
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3   Mean   :0.1485  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000  
    restecg          thalach          exang           oldpeak         slope             ca        
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :1.0000   Median :153.0   Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
 Mean   :0.5281   Mean   :149.6   Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
 3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
      thal           target      
 Min.   :0.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.0000  
 Median :2.000   Median :1.0000  
 Mean   :2.314   Mean   :0.5446  
 3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.0000  
```

Figure A.3: Structure of the data before cleaning



```
spec_tbl_df [303 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ age     : num [1:303] 63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : num [1:303] 1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : num [1:303] 3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: num [1:303] 145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : num [1:303] 1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : num [1:303] 0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : num [1:303] 0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : num [1:303] 0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : num [1:303] 1 2 2 2 2 1 2 3 3 2 ...
 $ target  : num [1:303] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   age = col_double(),
  ..   sex = col_double(),
  ..   cp = col_double(),
  ..   trestbps = col_double(),
  ..   chol = col_double(),
  ..   fbs = col_double(),
  ..   restecg = col_double(),
  ..   thalach = col_double(),
  ..   exang = col_double(),
  ..   oldpeak = col_double(),
  ..   slope = col_double(),
  ..   ca = col_double(),
  ..   thal = col_double(),
  ..   target = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Figure A.4: Structure of the data after cleaning



```
tibble [302 × 14] (S3: tbl_df/tbl/data.frame)
 $ age     : Factor w/ 8 levels "teens","twenties",..: 6 3 4 5 5 5 5 4 5 5 ...
 $ sex     : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 1 2 2 2 ...
 $ cp      : Factor w/ 4 levels "typical_angina",..: 4 3 2 2 1 1 2 2 3 3 ...
 $ trestbps: Factor w/ 5 levels "optimal","prehypertension",..: 3 2 2 1 1 2 2 1 4 3 ...
 $ chol    : Factor w/ 3 levels "Healthy Chol",..: 2 3 2 2 3 1 3 3 1 1 ...
 $ fbs     : Factor w/ 2 levels "Less than 120mg/dl fasting blood sugar",..: 2 1 1 1 1 1 1 1 2 1 ...
 $ restecg : Factor w/ 3 levels "Normal","ST-T_wave_abnormality",..: 1 2 1 2 2 2 1 2 2 2 ...
 $ thalach : Factor w/ 3 levels "low","normal",..: 2 3 3 3 3 2 3 3 3 3 ...
 $ exang   : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...
 $ oldpeak : chr [1:302] "More than Zero" "More than Zero" "More than Zero" "More than Zero" ...
 $ slope   : Factor w/ 3 levels "upsloping","flat",..: 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 "absence","normal",..: 2 3 3 3 3 2 3 4 4 3 ...
 $ target  : Factor w/ 2 levels "low risk","high risk": 2 2 2 2 2 2 2 2 2 2 ...
```

Figure A.5: Distribution of age

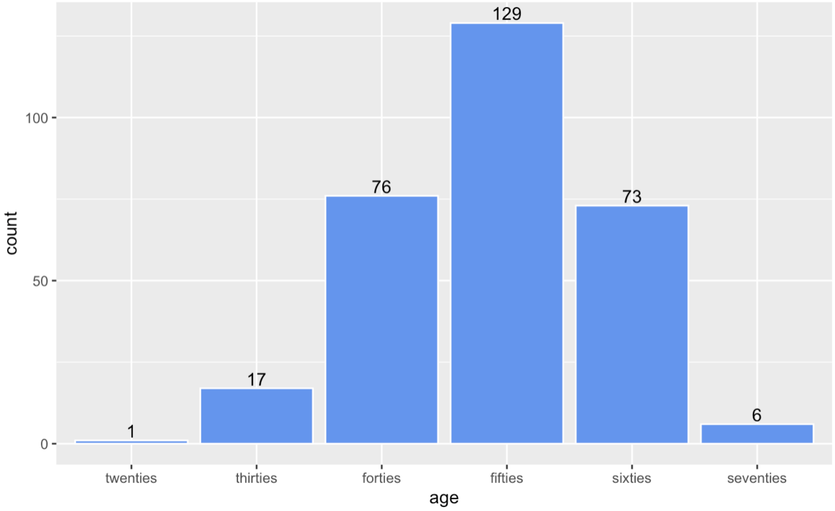


Figure A.6: Distribution of sex

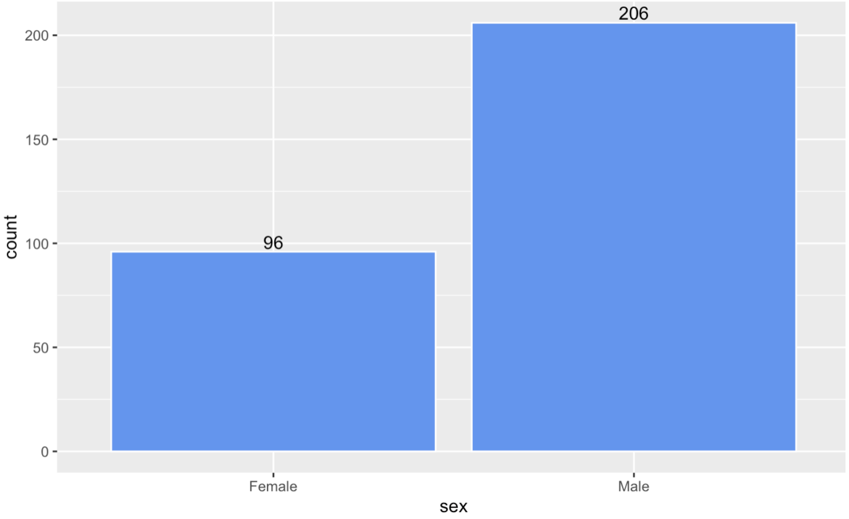


Figure A.7: Distribution of thal (evidence for blood disease)

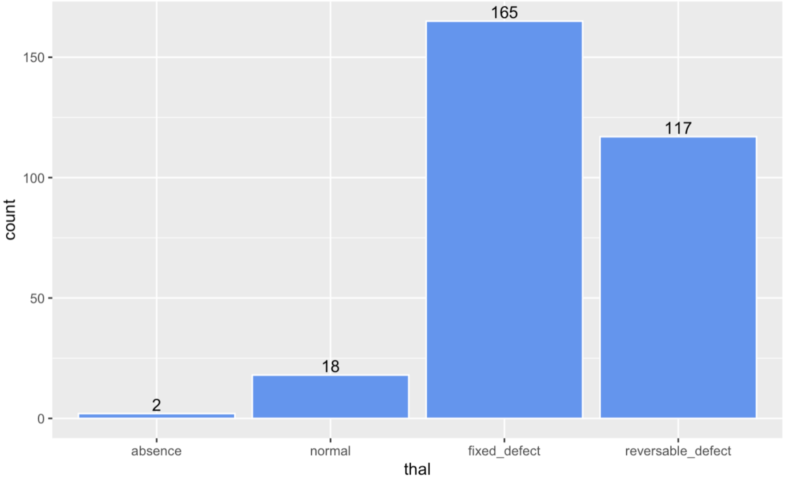


Figure A.8: Distribution of target (high risk or low risk)

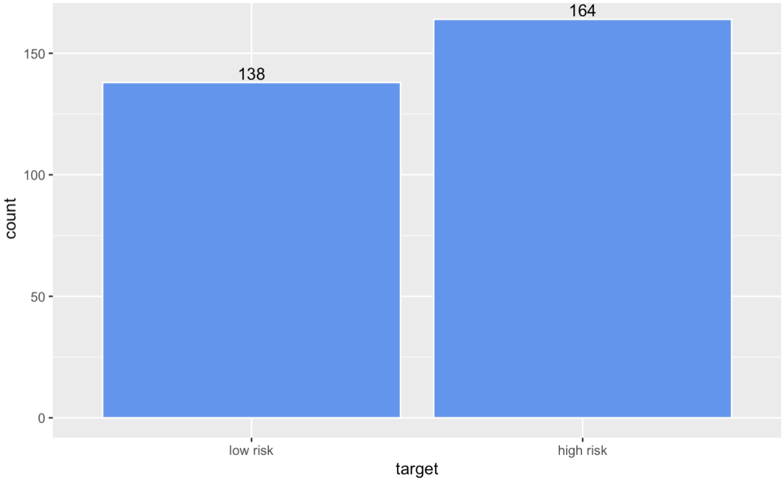


Figure A.9: Distribution of cp

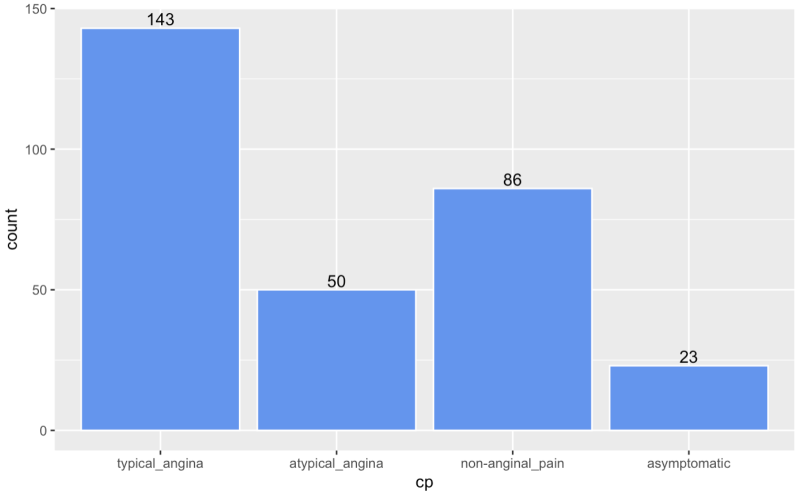


Figure A.10: Distribution of trestbps

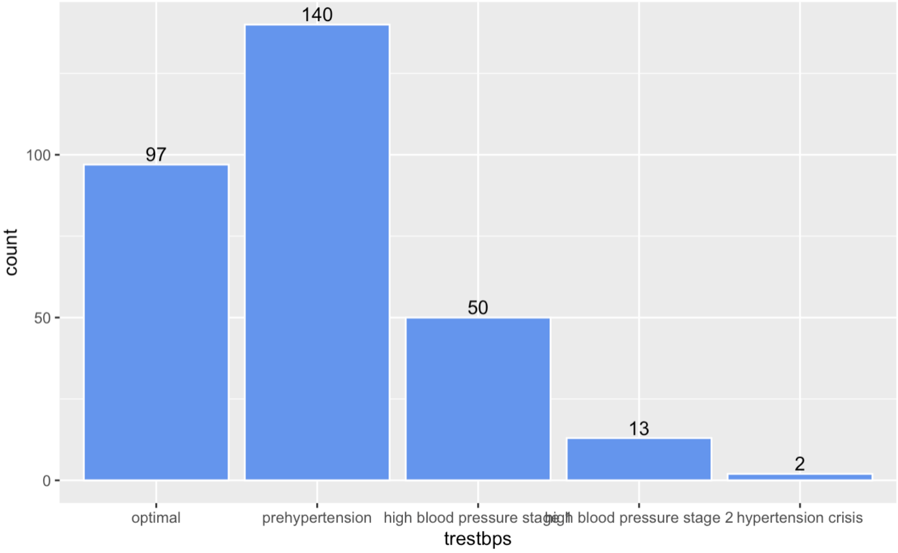


Figure A.11: Distribution of cholesterol

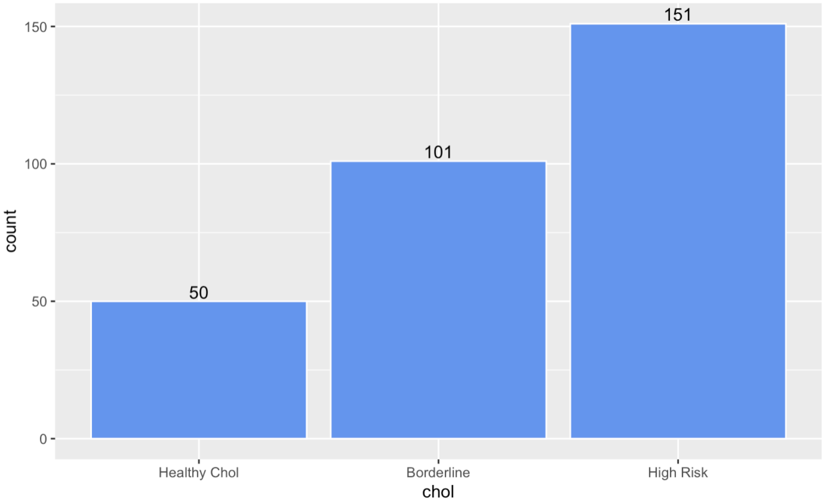


Figure A.12: Distribution of fasting blood sugar

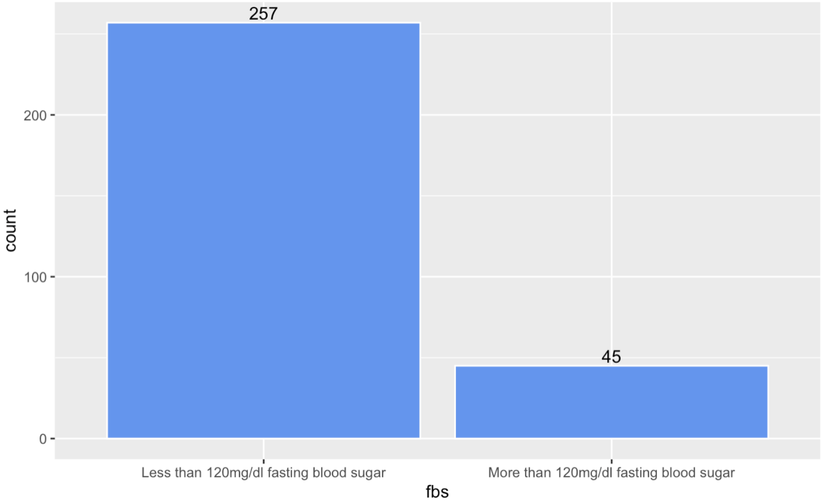


Figure A.13: Distribution of resting electrocardiographic results

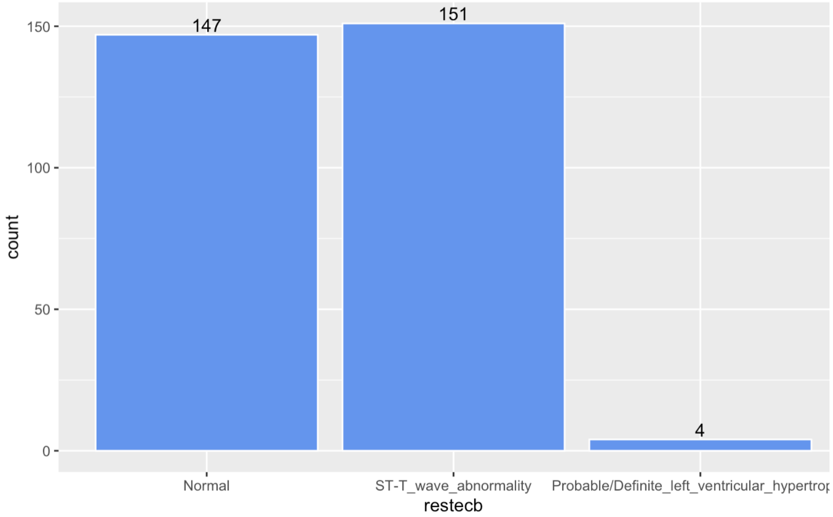


Figure A.14: Distribution of thalach

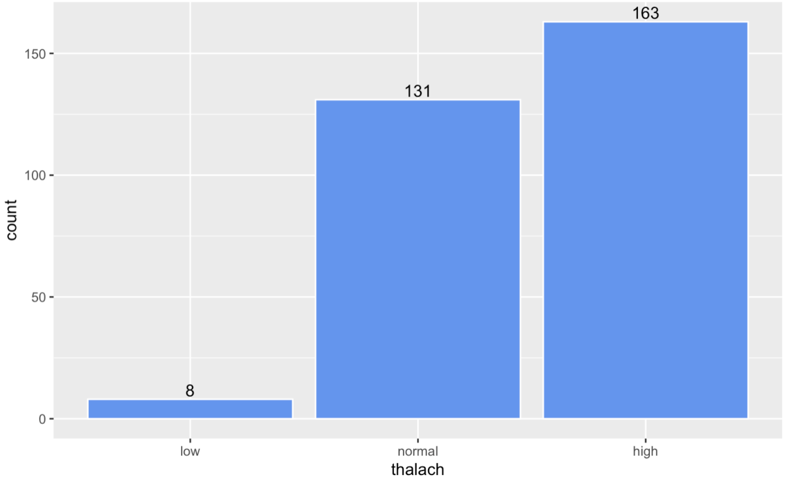


Figure A.15: Distribution of exang

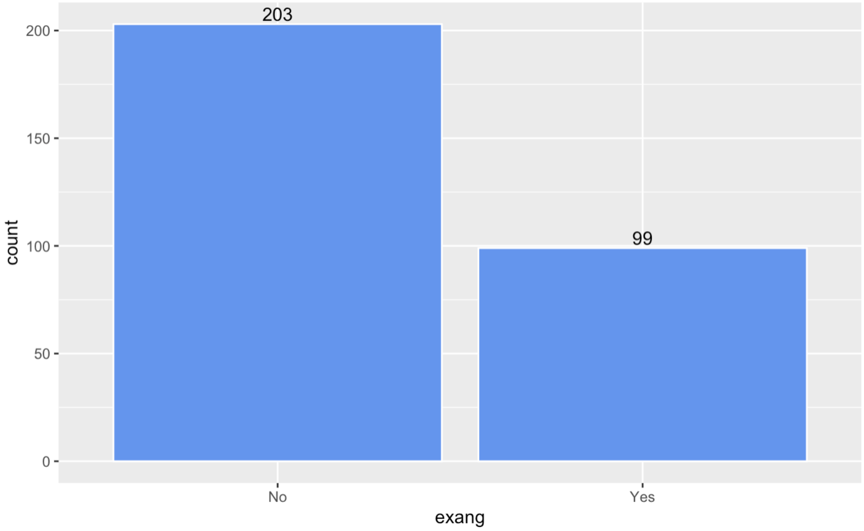


Figure A.16: Distribution of oldpeak

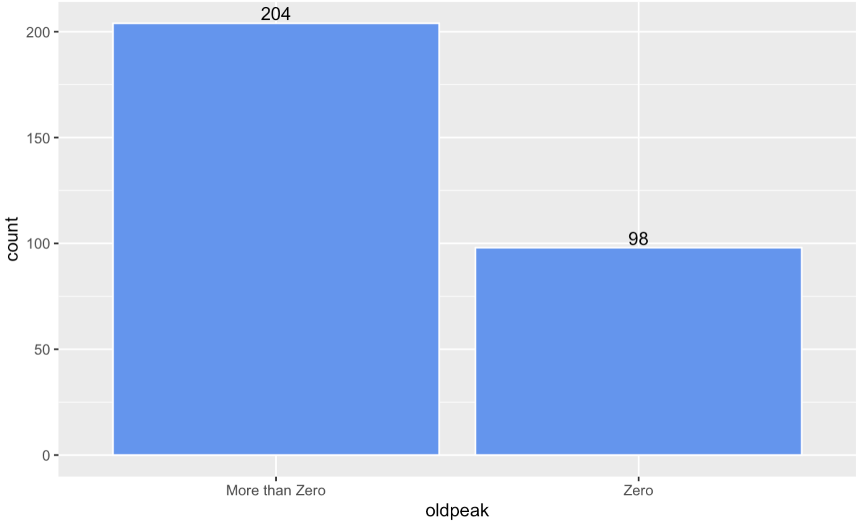


Figure A.17:

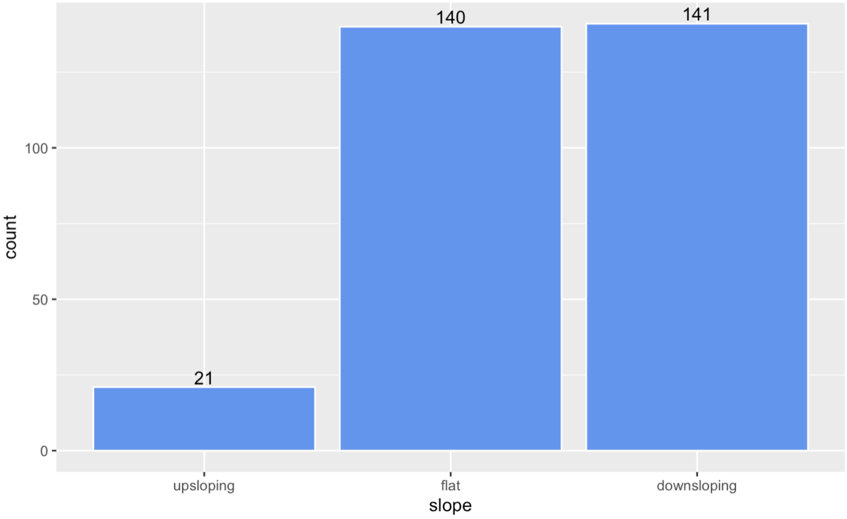


Figure A.18

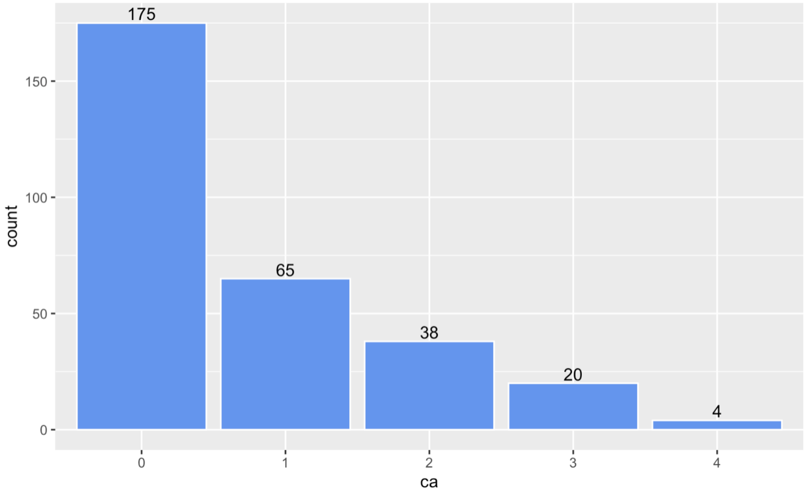


Figure A.19

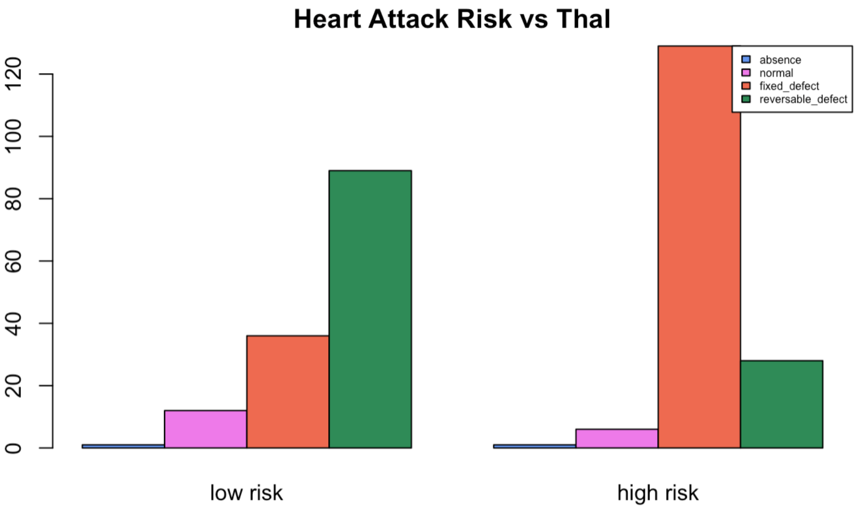


Figure A.20

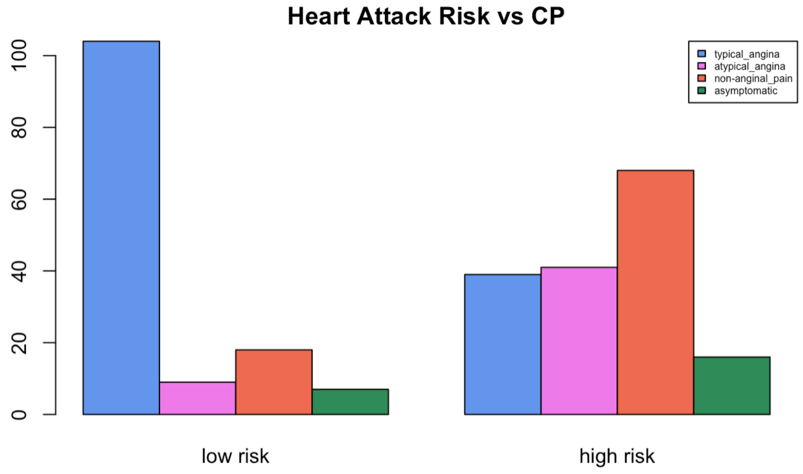


Figure A.21

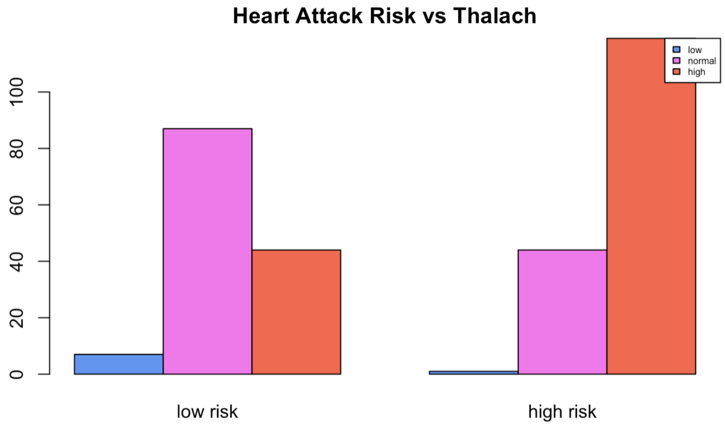


Figure A.22

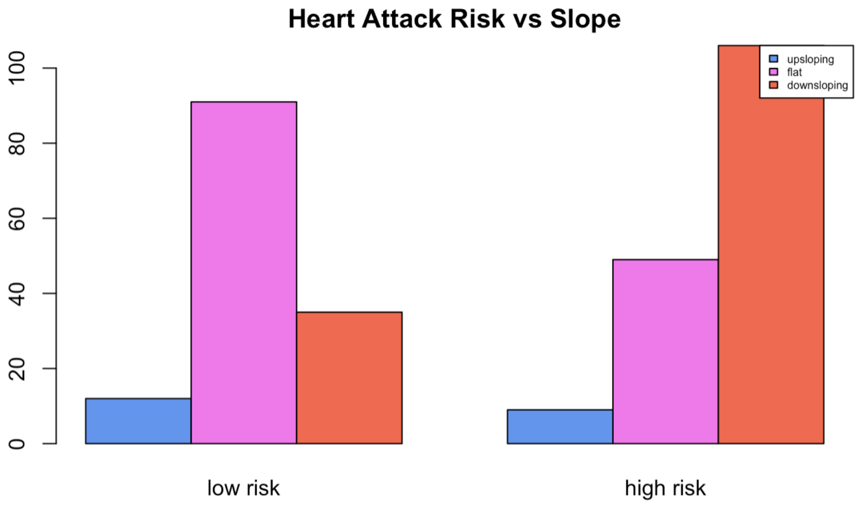


Figure A.23

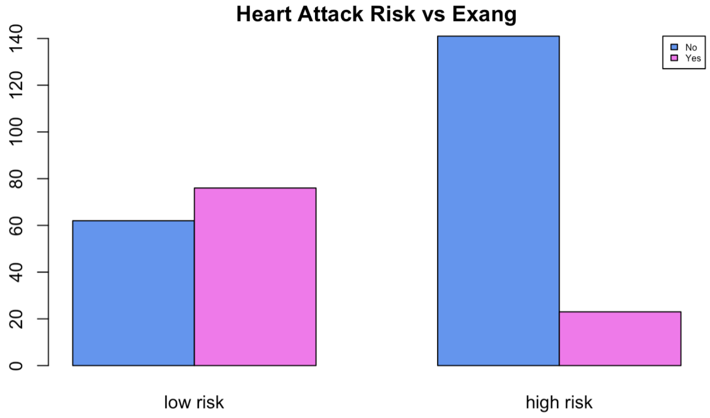


Figure A.24

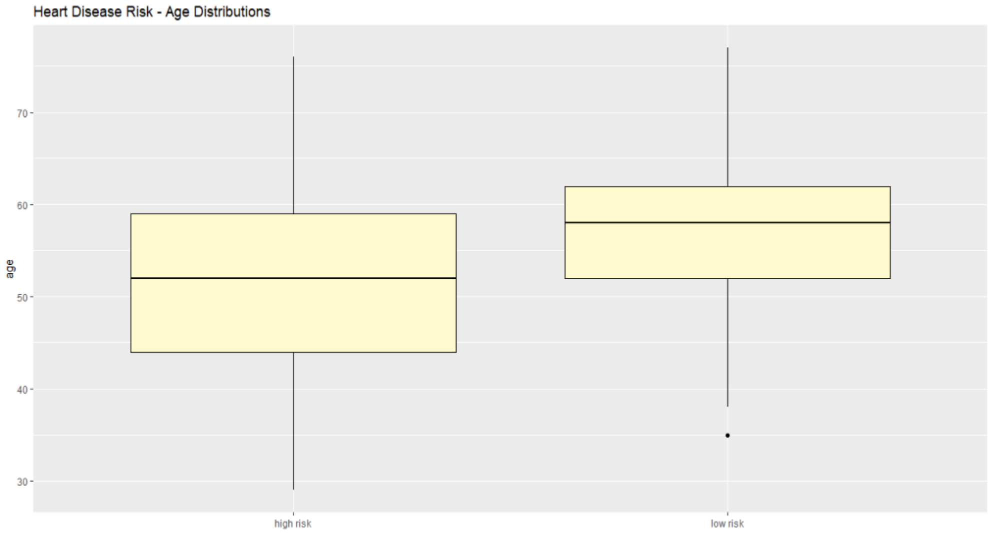


Figure A.25

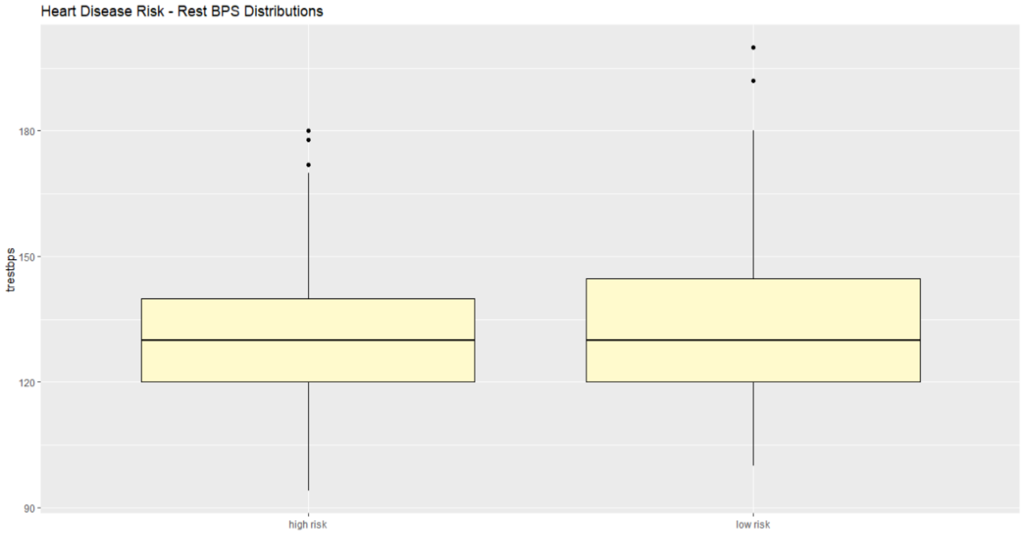


Figure A.26

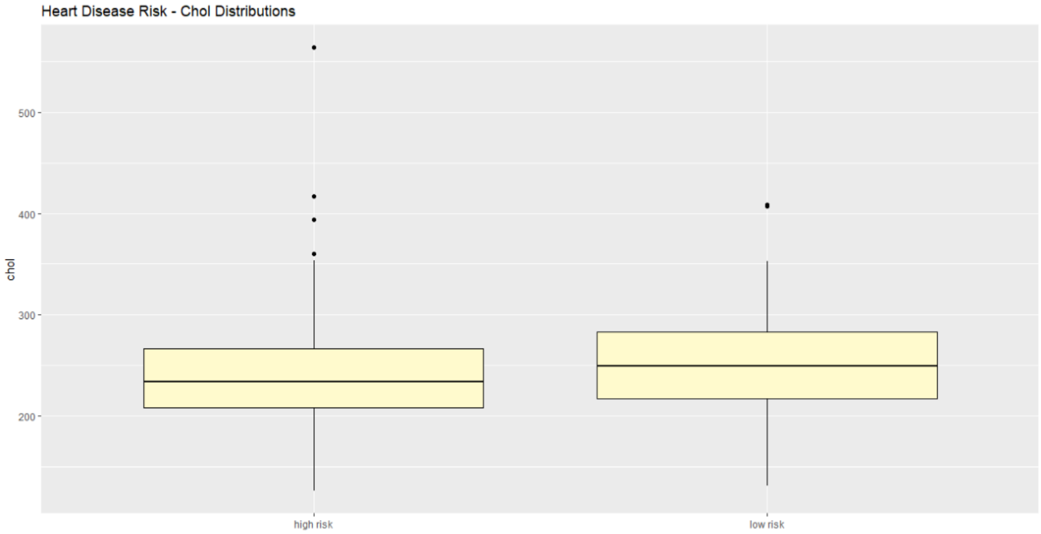


Figure A.27

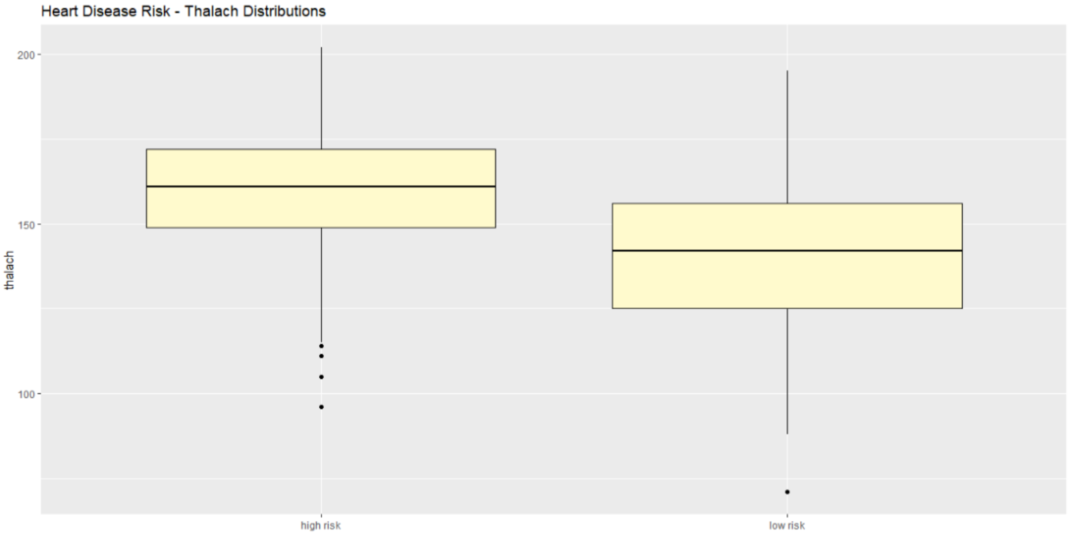


Figure A.28

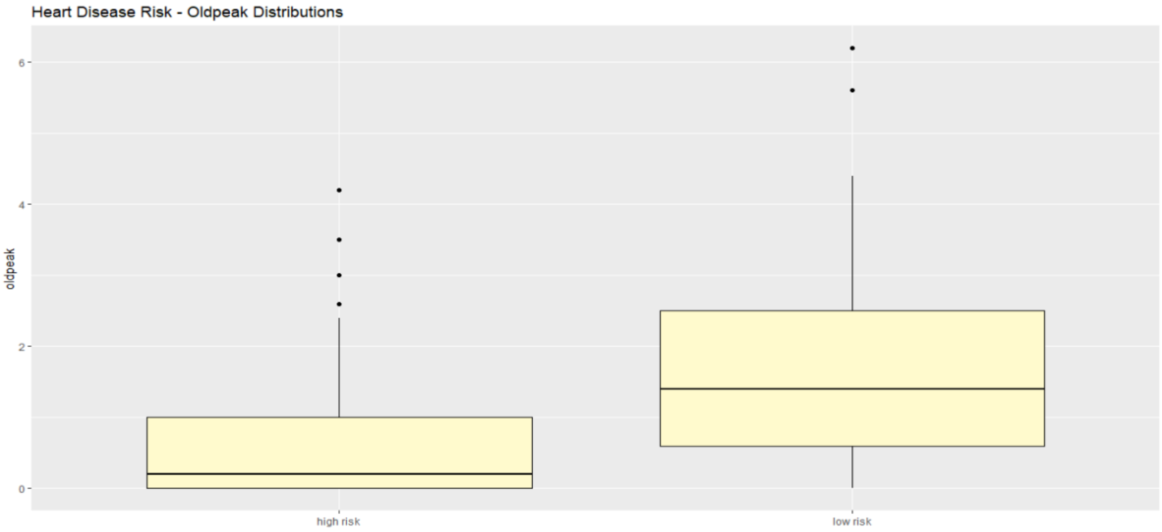
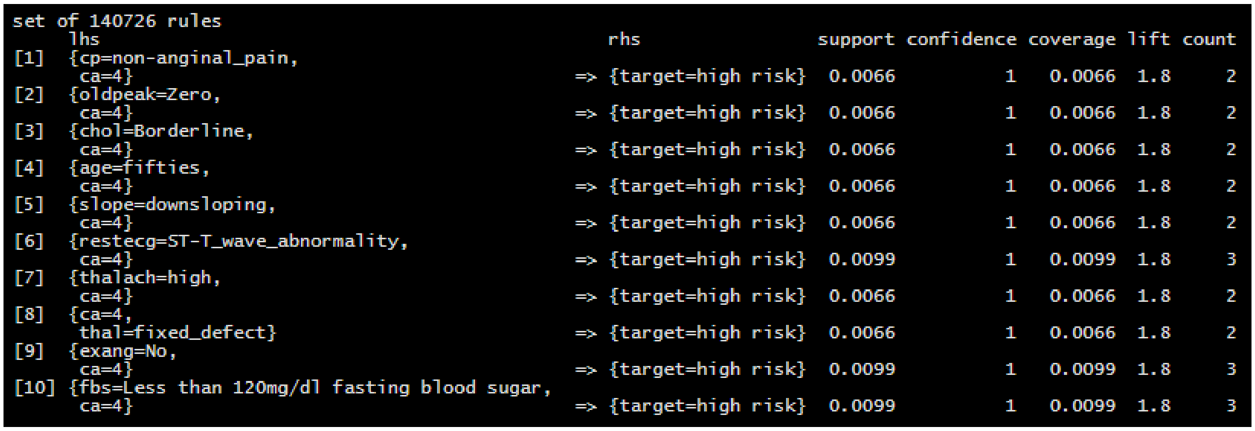
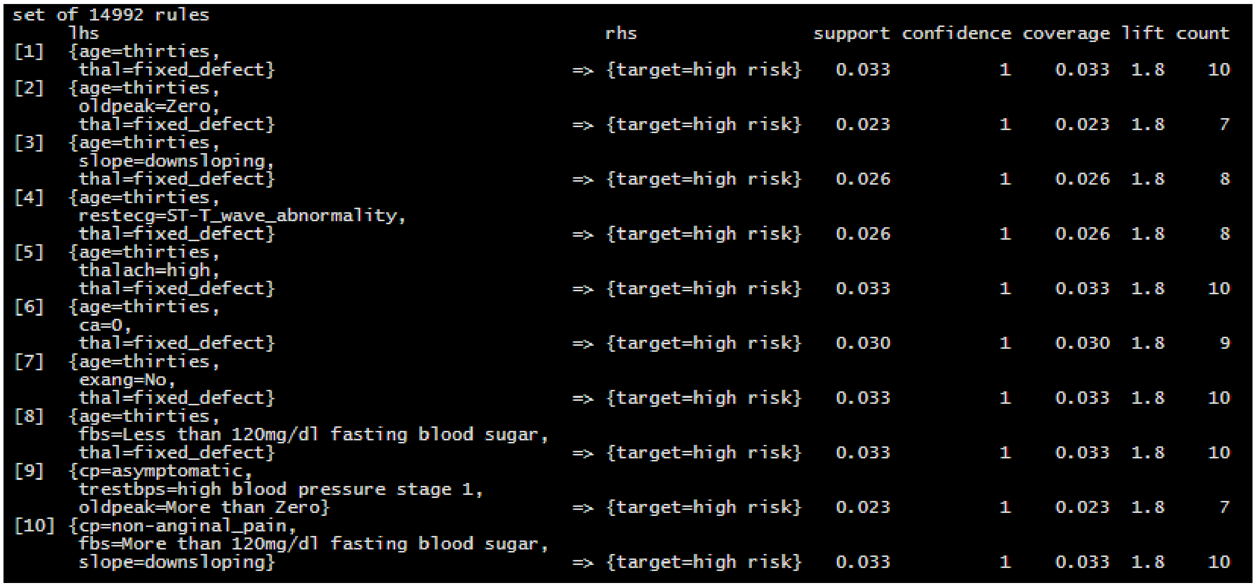


Figure B.1: ARM model with minimum support of .002, confidence of .8, and minimum length of 3 items. Top 10 rules displayed, ordered by lift descending, out of 140,726 rules.



```
set of 140726 rules
     lhs                                                rhs                support confidence coverage lift count
[1]  {cp=non-anginal_pain,
      ca=4}                                          => {target=high risk}  0.0066          1   0.0066  1.8     2
[2]  {oldpeak=Zero,
      ca=4}                                          => {target=high risk}  0.0066          1   0.0066  1.8     2
[3]  {chol=Borderline,
      ca=4}                                          => {target=high risk}  0.0066          1   0.0066  1.8     2
[4]  {age=fifties,
      ca=4}                                          => {target=high risk}  0.0066          1   0.0066  1.8     2
[5]  {slope=downsloping,
      ca=4}                                          => {target=high risk}  0.0066          1   0.0066  1.8     2
[6]  {restecg=ST-T_wave_abnormality,
      ca=4}                                          => {target=high risk}  0.0099          1   0.0099  1.8     3
[7]  {thalach=high,
      ca=4}                                          => {target=high risk}  0.0066          1   0.0066  1.8     2
[8]  {ca=4,
      thal=fixed_defect}                             => {target=high risk}  0.0066          1   0.0066  1.8     2
[9]  {exang=No,
      ca=4}                                          => {target=high risk}  0.0099          1   0.0099  1.8     3
[10] {fbs=Less than 120mg/dl fasting blood sugar,
      ca=4}                                          => {target=high risk}  0.0099          1   0.0099  1.8     3
```

Figure B.2: ARM model with minimum support of .02, confidence of .85, and minimum length of 3 items. Top 10 rules displayed, ordered by lift descending, out of 14,992 rules.



```
set of 14992 rules
     lhs                                                rhs                support confidence coverage lift count
[1]  {age=thirties,
      thal=fixed_defect}                             => {target=high risk}   0.033          1    0.033  1.8    10
[2]  {age=thirties,
      oldpeak=Zero,
      thal=fixed_defect}                             => {target=high risk}   0.023          1    0.023  1.8     7
[3]  {age=thirties,
      slope=downsloping,
      thal=fixed_defect}                             => {target=high risk}   0.026          1    0.026  1.8     8
[4]  {age=thirties,
      restecg=ST-T_wave_abnormality,
      thal=fixed_defect}                             => {target=high risk}   0.026          1    0.026  1.8     8
[5]  {age=thirties,
      thalach=high,
      thal=fixed_defect}                             => {target=high risk}   0.033          1    0.033  1.8    10
[6]  {age=thirties,
      ca=0,
      thal=fixed_defect}                             => {target=high risk}   0.030          1    0.030  1.8     9
[7]  {age=thirties,
      exang=No,
      thal=fixed_defect}                             => {target=high risk}   0.033          1    0.033  1.8    10
[8]  {age=thirties,
      fbs=Less than 120mg/dl fasting blood sugar,
      thal=fixed_defect}                             => {target=high risk}   0.033          1    0.033  1.8    10
[9]  {cp=asymptomatic,
      trestbps=high blood pressure stage 1,
      oldpeak=More than Zero}                        => {target=high risk}   0.023          1    0.023  1.8     7
[10] {cp=non-anginal_pain,
      fbs=More than 120mg/dl fasting blood sugar,
      slope=downsloping}                             => {target=high risk}   0.033          1    0.033  1.8    10
```