IST707 Data Analytics | Section M402

**Predicting Heart Disease**

Jacqueline Parks, Laura Larregui, Courtney Rowland, Jacob Arluck

horizontal line

# 

# Introduction

According to the Centers for Disease Control and Prevention, heart disease is the term given to a range of heart condition types. For instance, coronary artery disease or CAD, where a buildup of plaque in the arteries affects blood flow to the heart is the most prevalent type of heart disease in the United States. Heart attacks are another type of heart disease, which is the event caused by blocked arteries where the heart tissue begins to die from lack of oxygen it receives through the blood. These diseases are often referred to as being silent because they sometimes do not get diagnosed until the person experiences an event like a heart attack or arrhythmia. Some common symptoms of a heart attack are chest pain/discomfort, heartburn, upper back or neck pain, nausea, extreme fatigue, dizziness, and shortness of breath.

Some statistics about heart disease include it being the leading cause of death for both men and women of most racial and ethnic groups within the U.S. Every 37 seconds a person in the United States dies from cardiovascular disease, approximately 647,000 each year. In regard to expenditure, it costs the U.S. about $219 billion each year to cover health care services, medicine, and productivity loss caused by deaths in the years 2014-15.

Based on this information, heart disease is a major cause for concern. Due to a large number of risk factors such as high blood pressure, cholesterol levels, smoking, diabetes, obesity, diet, physical inactivity, and excessive alcohol usage, it can be difficult to identify heart disease in individuals. Now as data science is shaping the future, scientists are turning to data analysts skilled in machine learning techniques to generate algorithms for predicting heart disease.

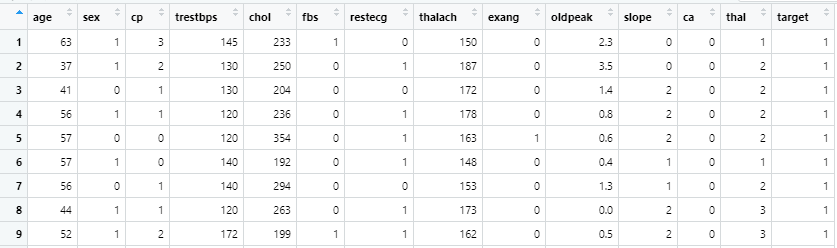
# Analysis

## The Data

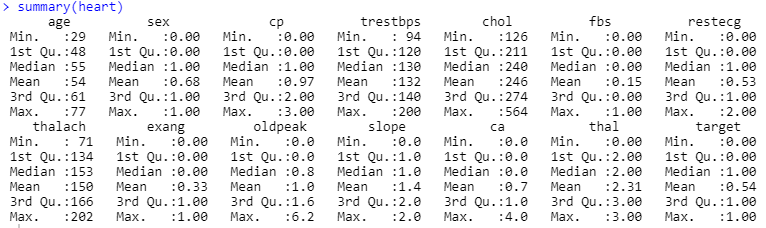
Our data is a heart disease data set from Kaggle but originated from the University of California, Irvine’s (UCI) machine learning repository. It is compiled from four databases which are based on data collected in Cleveland, Hungary, Switzerland, and Virginia Long Beach. The data set contains 76 attributes but the project focuses on a subset of the 14 most relevant attributes. There are a total of 303 instances formatted in a CSV file.

The fourteen variables included in the data set are: (1) age (continuous) , (2) gender (0=female, 1=male), (3) whether the patient has had any chest pains (0: Typical angina, 1: Atypical angina, 2: Non-anginal pain, 3: Asymptomatic), (4) their resting blood pressure on hospital admission (continuous), (5) their cholesterol (continuous), (6) whether they have a fasting blood sugar over 120 (which would indicate diabetes; marked as true or false), (7) whether they have a normal or irregular ECG ( normal 0, Level 1, Level 2), (8) their maximum heart rate during an exercise stress test (continuous), (9) whether they exhibit exercise-induced angina (chest pain; 0=no, 1=yes), (10) the peak of the ST depression caused by exercise (continuous), (11) the slope of the peak exercise ST segment (ST segment depression can indicate myocardial ischemia; 0=up, 1=down, 2=flat), (12) the number of major vessels colored by fluoroscopy (meaning containing disease; 0-3), (13) evidence of disease in a thallium test (0=normal, 1=fixed defect, 2=reversible defect), and (14) our target variable (presence of heart disease; 0=higher chance, 1=lower chance).

**Figure 1. Sample view of the heart dataset in R.**

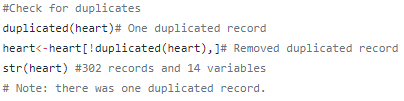


**Image 1. A statistical summary for each of the 14 variables, target being the goal variable.**



**Preprocessing**

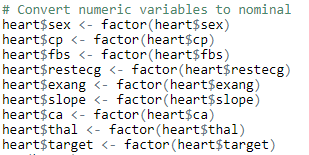
1. The original 14 variables from the Kaggle data set were kept since they were the most commonly experimented variables mentioned on the original website.
2. There were no NAs but one duplicate record that was deleted.



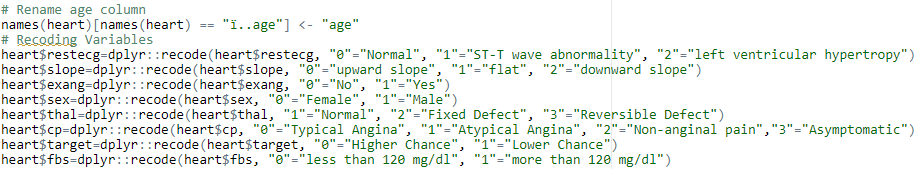
1. For thal, the dataset had 0,1,2,3 values. However, the data set metadata did not identify what the value 0 means, so two records were removed from the dataset.



1. Numeric variables were converted to nominal

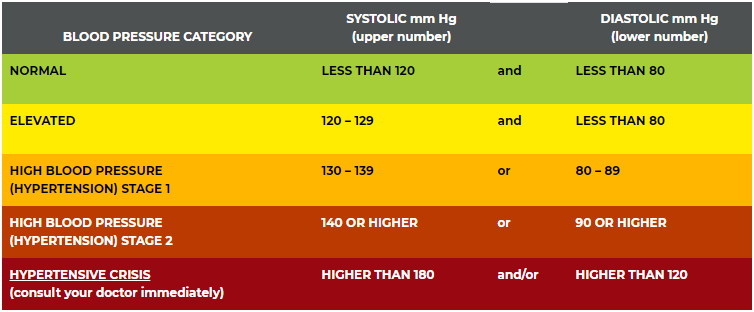


1. Age was renamed and eight variables were re-coded using the dplyr package.

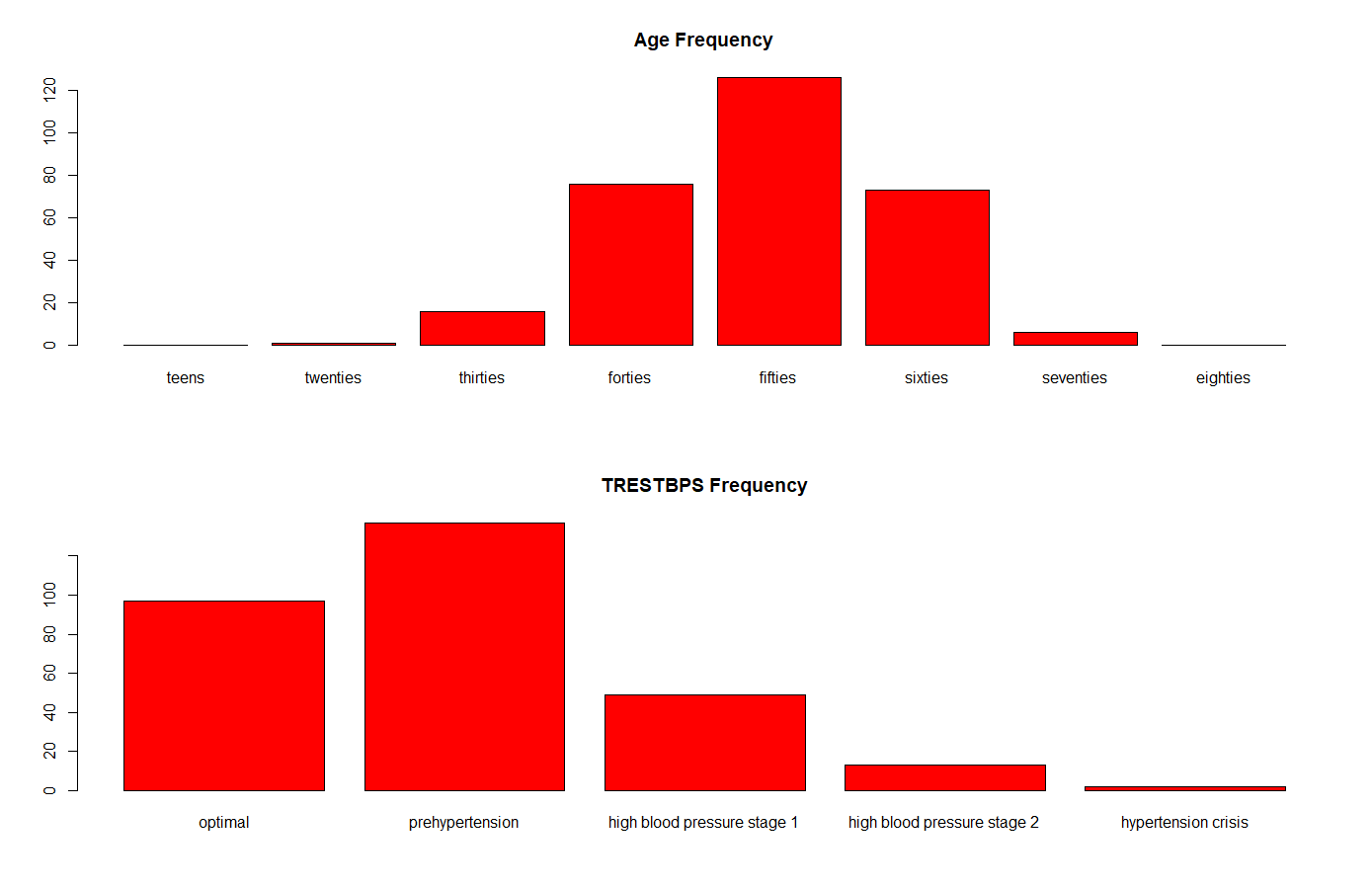


1. The variables age, chol, and trestbps were discretized as seen in figure 2 and image 3.

Image 2. Chart used when discretizing trestbps variable taken from the American Heart Association website.



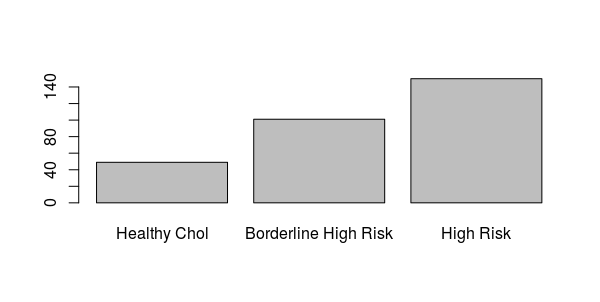
**Figure 2. Plots displaying the preprocessed discretization of variables Age and trestbps, with resting blood pressure following the American Heart Association guidelines provided in image 2.**



**Discretizing chol**

Research was necessary to determine the ranges to discretize cholesterol levels accordingly. According to statinusage.com, since high cholesterol is considered to be a silent condition since it shows no symptoms, a series of blood tests called lipid profile are necessary to measure these levels. Measurement is in milligrams per deciliter of blood--or mg/dL. Cholesterol levels between 200-240 mg/dL are considered to be “borderline high risk” and levels beyond 240 mg/dL are “high risk” for heart disease.

**Image 3. Plot displaying chol variable after discretization.**



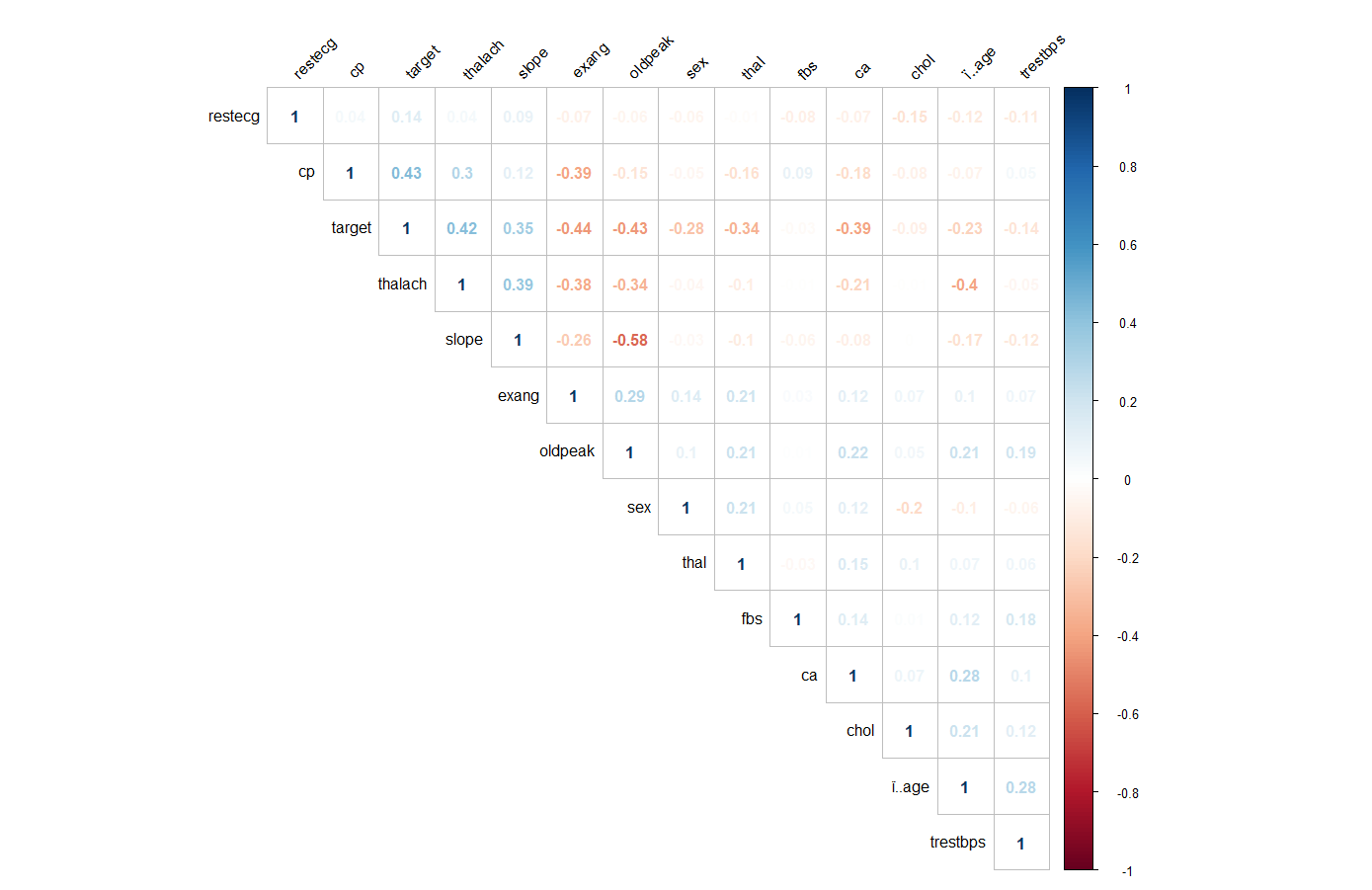
In order to perform data mining methods, the dataset was further transformed by splitting into two subsets (one for training and the other one for testing). The training subset contains 66% random records of the original dataset, while the testing subset has the remaining 33%.

## 

## EDA: Statistical and Visual

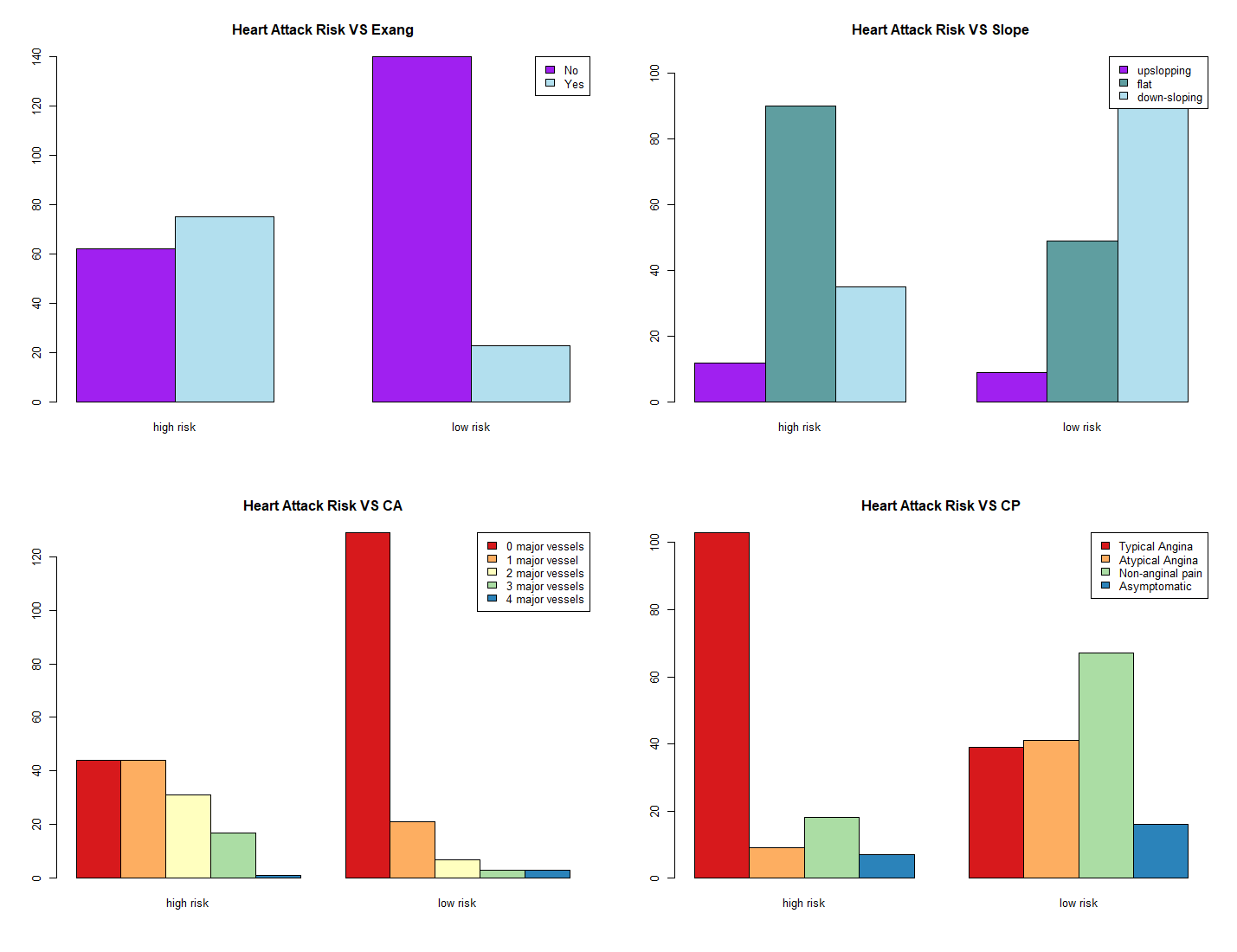
Before any data mining methods were performed, an exploratory data analysis was carried out. First, a correlation matrix was made to observe which attributes correlated with the target variable.

**Figure 3. Correlation Matrix**



From the correlation matrix, one can see that the variables cp (0.43), thalach (0.42), and slope (0.35) directly correlated with the target variable, whereas exang (-0.44), oldpeak (-0.43), and ca (-0.39) negatively correlated with target. From these observations, barplots were made from the categorical variables with significant correlations.

**Figure 4. Barplots of the categorical variables with significant correlations**



From the barplots, one can presume that people with a high risk of heart disease have 0 to 1 major vessels colored by fluoroscopy, a flat slope of the peak exercise ST segment, have exercise-induced angina, and have typical angina chest pain.

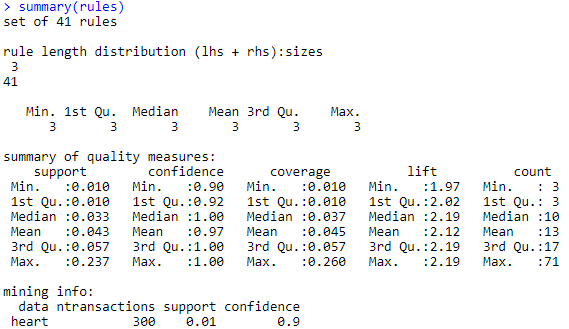
## Methods

### Association Rule Mining

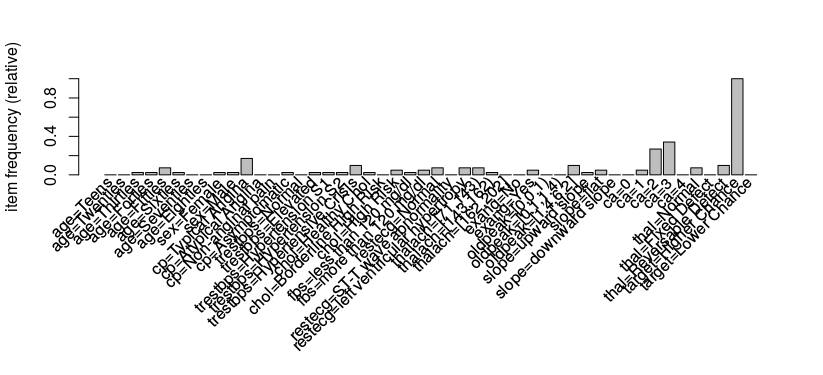
ARM, as the name implies, is a procedure that looks for patterns in data sets. ARM is appropriate for categorical (nominal) data. Hence, the reason why all the variables in the data set were changed from numerical to nominal with the exception of chol, thalach, and oldpeak.

ARM has three parameters: support, confidence, and lift. Support is how frequently a relationship appears in the data set. Confidence is the number of times these relationships have been found to be true. Lastly, the lift is a correlation between the left-hand side and the right-hand side. Thus, the left-hand side should be greater than 1 for the rules to be meaningful. The lift measurement is significant since the confidence measure can be sometimes misleading.

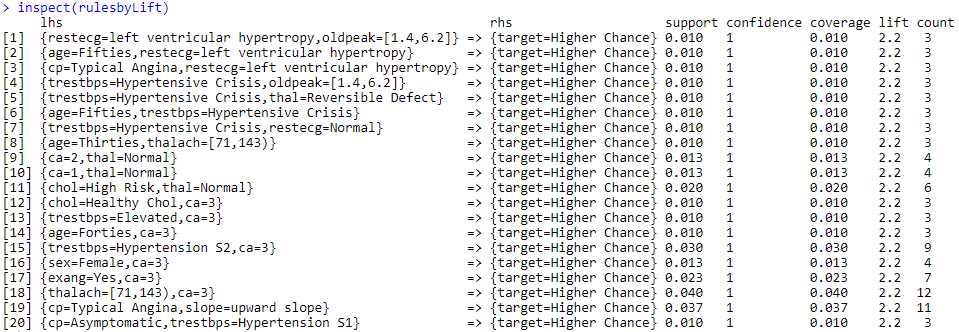
**Image 4. Overview of mined rules. The apriori function was used to find all the rules, with minimum support of 1% and a confidence of 0.9. The result of the mining algorithm is a set of 41 rules.**



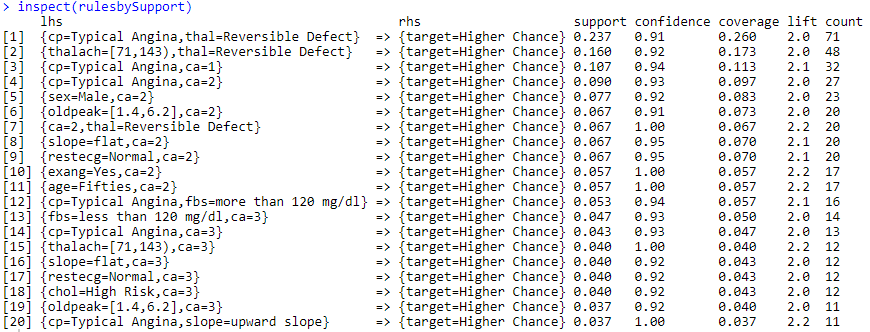
**Figure 5. Item frequencies of items in the heart data set with support greater than 10%.**



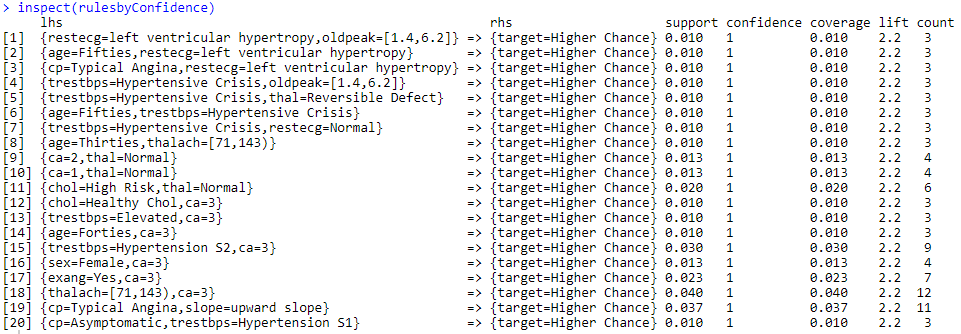
**Image 5. Screenshot of top 20 rules sorted by Lift.**



**Image 6. Screenshot of top 20 rules sorted by Support.**



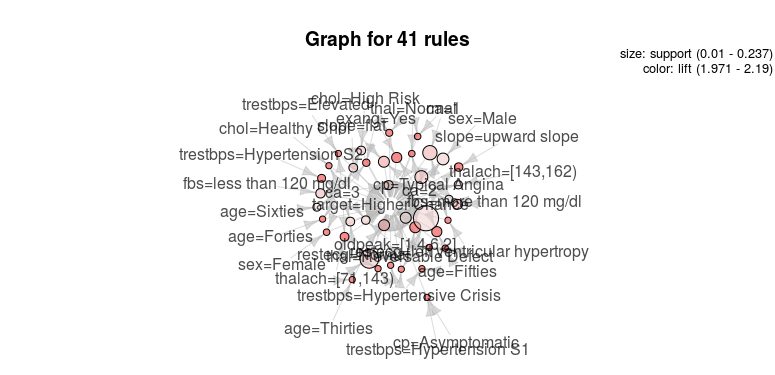
**Image 7. Screenshot of top 20 rules sorted by Confidence.**



**Table 1: The top 5 most interesting rules mined.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Rule:** | **LHS** | **RHS** | **Support** | **Confidence** | **Coverage** | **Lift** | **Count** |
| 1 | age=Fifties, trestbps= Hypertensive Crisis | Target = Higher Chance | 0.010 | 1 | 0.010 | 2.2 | 3 |
|  | Interpretation: Patients in their fifties admitted to the hospital with resting blood pressure in a hypertensive crisis. | | | | | | |
| 2 | thalach=[71,143), ca=3 | Target = Higher Chance | 0.040 | 1 | 0.040 | 2.2 | 12 |
|  | Interpretation: Patient had a maximum heart rate achieved between 71-143 and three major vessels containing disease. All three parameters were fairly high among the top sorted by confidence. | | | | | | |
| 3 | cp=Typical Angina, thal= Reversible Defect | Target = Higher Chance | 0.160 | 0.92 | 0.173 | 2.0 | 48 |
|  | Interpretation: Patients presenting with typical angina chest pain and had a heart status of reversible defect after the Thallium test. | | | | | | |
| 4 | sex=Male, ca=2 | Target = Higher Chance | 0.077 | 0.92 | 0.083 | 2.0 | 23 |
|  | Interpretation: Male patients possessing 2 diseased major vessels. | | | | | | |
| 5 | age=Thirties, thalach= [71,143) | Target = Higher Chance | 0.010 | 1 | 0.010 | 2.2 | 3 |
|  | Interpretation: A patient in their thirties with a maximum heart rate achieved between 71-143. This rule had a high confidence and high lift which was surprising to see even younger patients with healthy ranges still presenting a higher chance for heart disease. | | | | | | |

**Figure 5. Plot of total rules mined. Many descriptive variables can be seen visually that contribute to a target patient with high risk for heart disease such as those whose cholesterol level is considered ‘High Risk’, resting blood pressure in stage 1 hypertension, in the forties age range, etc.**



### Clustering and PCA

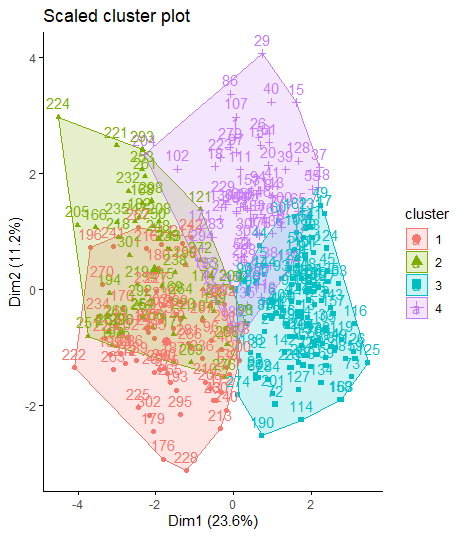
Cluster analysis is an unsupervised method (i.e. the data is not labeled) which seeks to discover if categories exist, and if they do, then it tries to identify what they are. In brief, the clustering method groups objects together based on their similarity and keeps apart the objects that are different. Thus, intracluster distances are minimized, and inter-cluster distances maximized.

There are two types of clustering: partitional and hierarchical. Partitional clustering sets a division of data objects into non-overlapping clusters such that each object is in exactly one subset. On the other hand, hierarchical clustering is a set of nested clusters organized in a hierarchical tree.

Clustering methods work best when data is normalized, preventing one variable from dominating the similarity metric. Accordingly, the data was separately preprocessed for clustering, keeping all variables numeric, and normalizing them such that standardized means were 0 and variances 1.

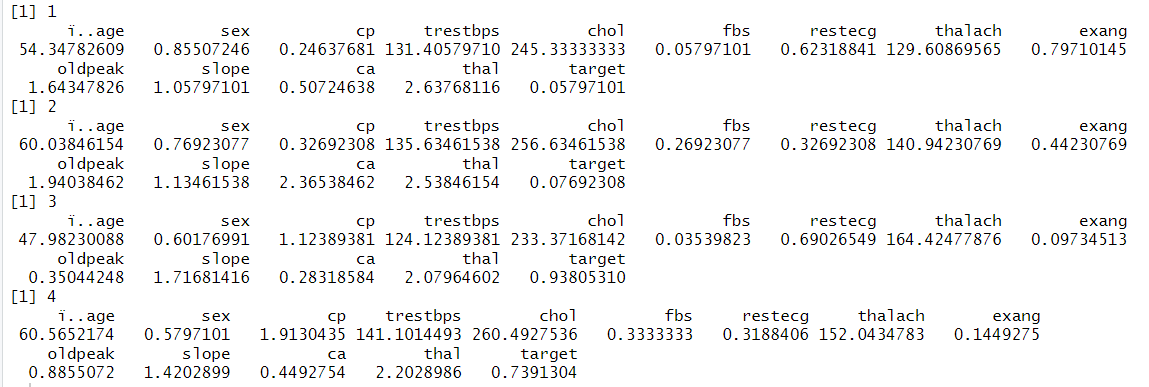
K-means partitional clustering was the most successful clustering technique. This technique works by picking some number of randomly-chosen centroids, and then assigning each point to the nearest centroid. The centroids are then recalculated to minimize distance, by setting them to the weighted average of all the points in the cluster. The process is then repeated, and continued until cluster identities are stable. Similar results were observed with 3, 4, or 5 initial clusters; as a representative example, 4 clusters were used.

The results of clustering can be most easily visualized by combining cluster analysis with principal components analysis (PCA). PCA is the most widely used dimensionality reduction technique. It breaks down a matrix of data points into a set of principal vectors and singular values, such that the vectors provide as “good as possible” an approximation of the original matrix when only a few vectors are used. A PCA plot shows data points according to their components along the two most significant (highest percentage of the total variance) vectors. Figure 6 displays the clustering overlaid on the PCA plot:



**Figure 6: 4-cluster k-means overlaid on a PCA chart. Analysis revealed that the x-axis is approximately a measure of overall heart disease risk, while the y-axis is roughly a measure of diabetes risk.**

Once clusters were assigned, the different characteristics of each cluster were analyzed by viewing means of variables for each cluster.



### Image 7.5: Screenshot of summary statistics by cluster.

This analysis revealed several interesting trends. Two clusters (3 and 4) were low-risk and two (1 and 2) high risk. Cluster 3 was more “typically” low risk - younger (average age 47 years), more female, lower blood sugar, lower cholesterol. Conversely, cluster 2 was “typically” high risk in terms of all of these metrics. The remaining two clusters were comparatively atypical in terms of what might be expected for a high- or low-risk patient. Perhaps these clusters represent patients who are genetically resistant to heart disease despite poor lifestyle choices (in the case of cluster 4) or genetically predisposed to heart disease (in the cast of cluster 1).

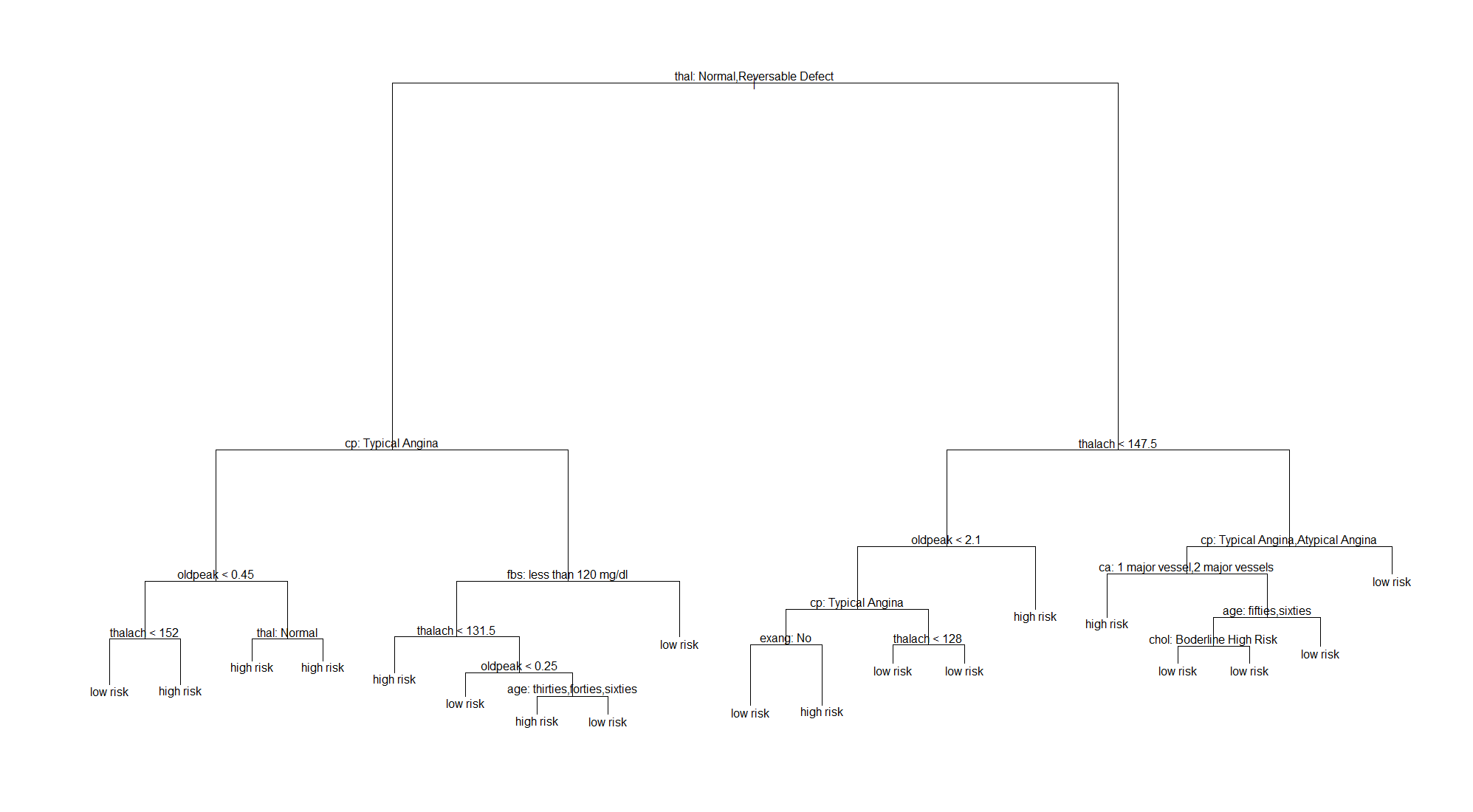
Additionally, inspection of thc clusters revealed a clear interpretation of the two most important dimensions found during PCA. The first axis corresponded very closely to overall heart disease risk; it suggests the existence of a “general factor of heart disease.” The second axis seems to indicate diabetes risk as distinct from heart disease risk. Of course, diabetes and heart disease are in fact related, but this analysis indicates that there is a component of diabetes risk that is orthogonal (independent) of heart disease risk.

### Decision Tree

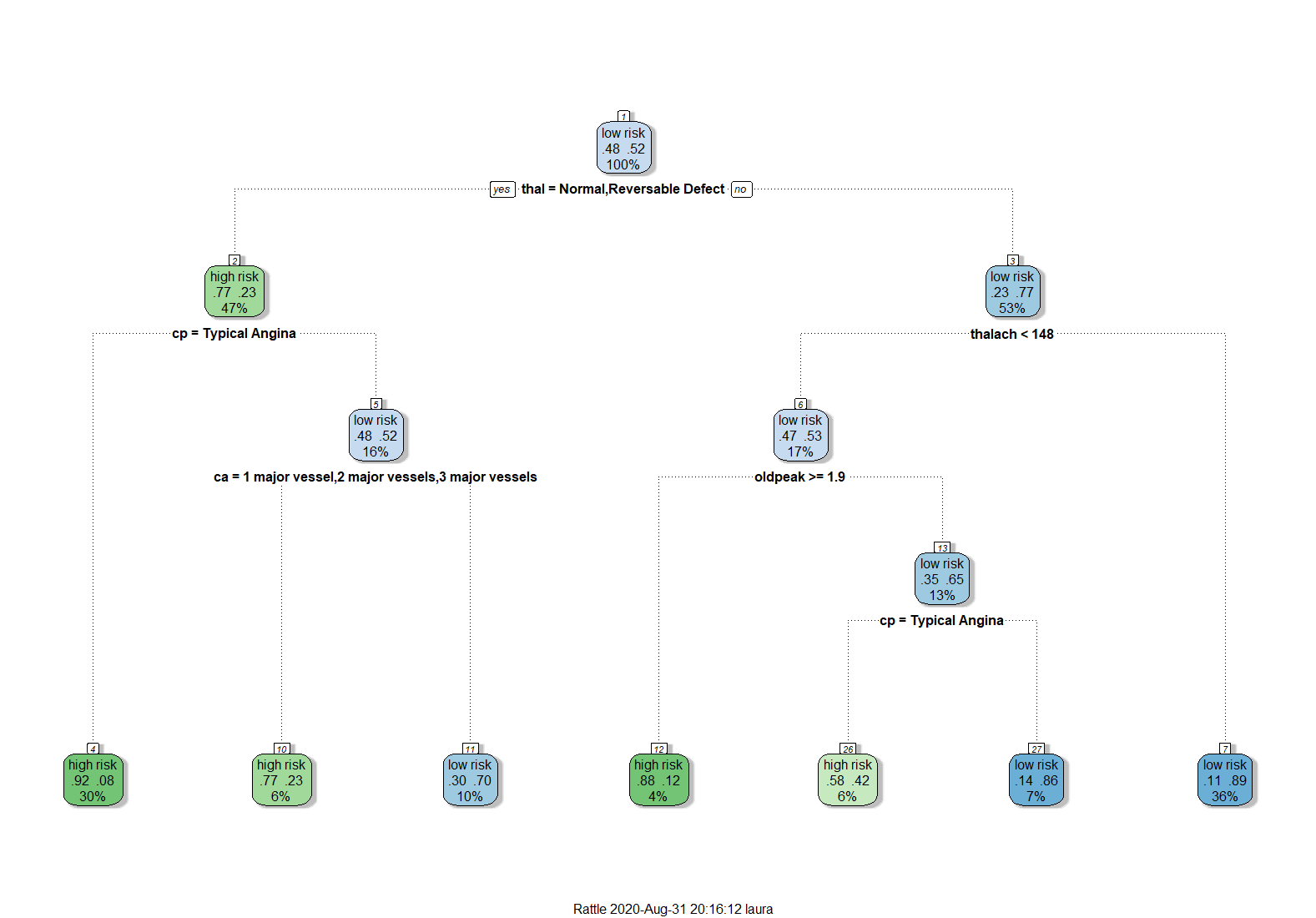
A decision tree diagram creates a prediction tree-like model to look for possible outcomes or events and their chances. Decision trees have two components, nodes, and leaves. Nodes examine a single variable and move to another node (the leaf). The first node is called a root. If a decision tree is composed of only the root, then the model is insignificant.

For this project, an initial decision tree model was made to observe which attributes were common in people with high heart disease risk. As one can see in **Figure 7**, the initial model was difficult to understand. Thus, a pruned tree model was needed to make further conclusions. Using the rpart function in R, an easier to understand model was elaborated (see **Figure 8**). This model had an accuracy of 80%.

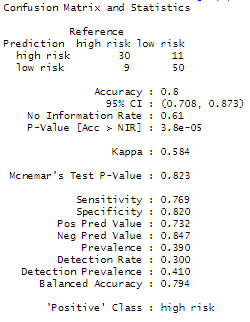
**Figure 7. Unpruned decision tree model**



**Figure 7. Pruned decision tree model**

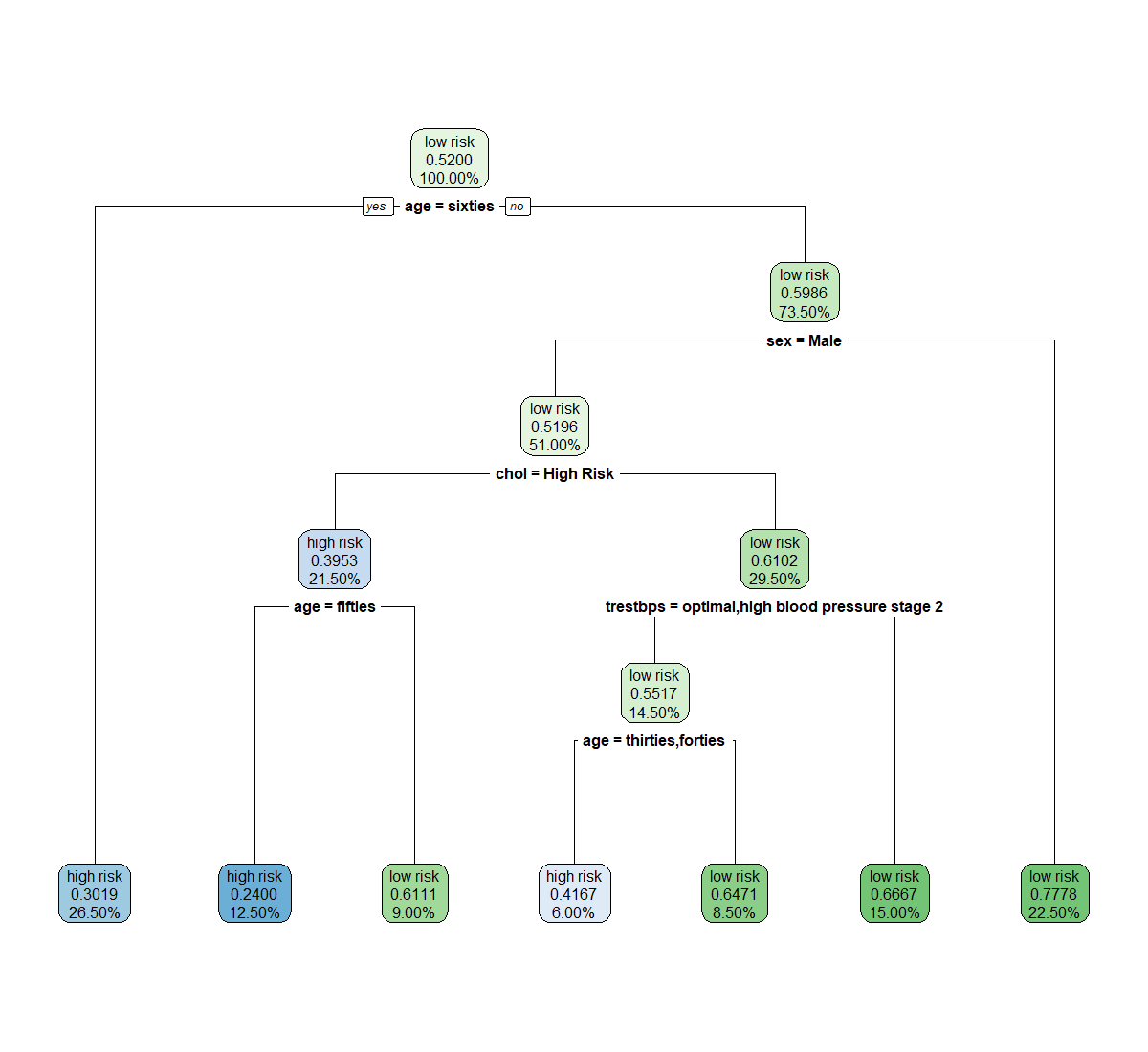


**Image 8. Pruned decision tree model confusion matrix**

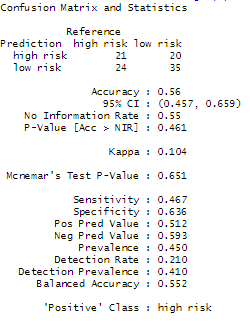


To provide further analysis, a decision tree model was made for the demographic variables in relation to their cholesterol and blood pressure and the variables with a high correlation. The accuracy achieved by these tree models was 56 % and 81% respectively.

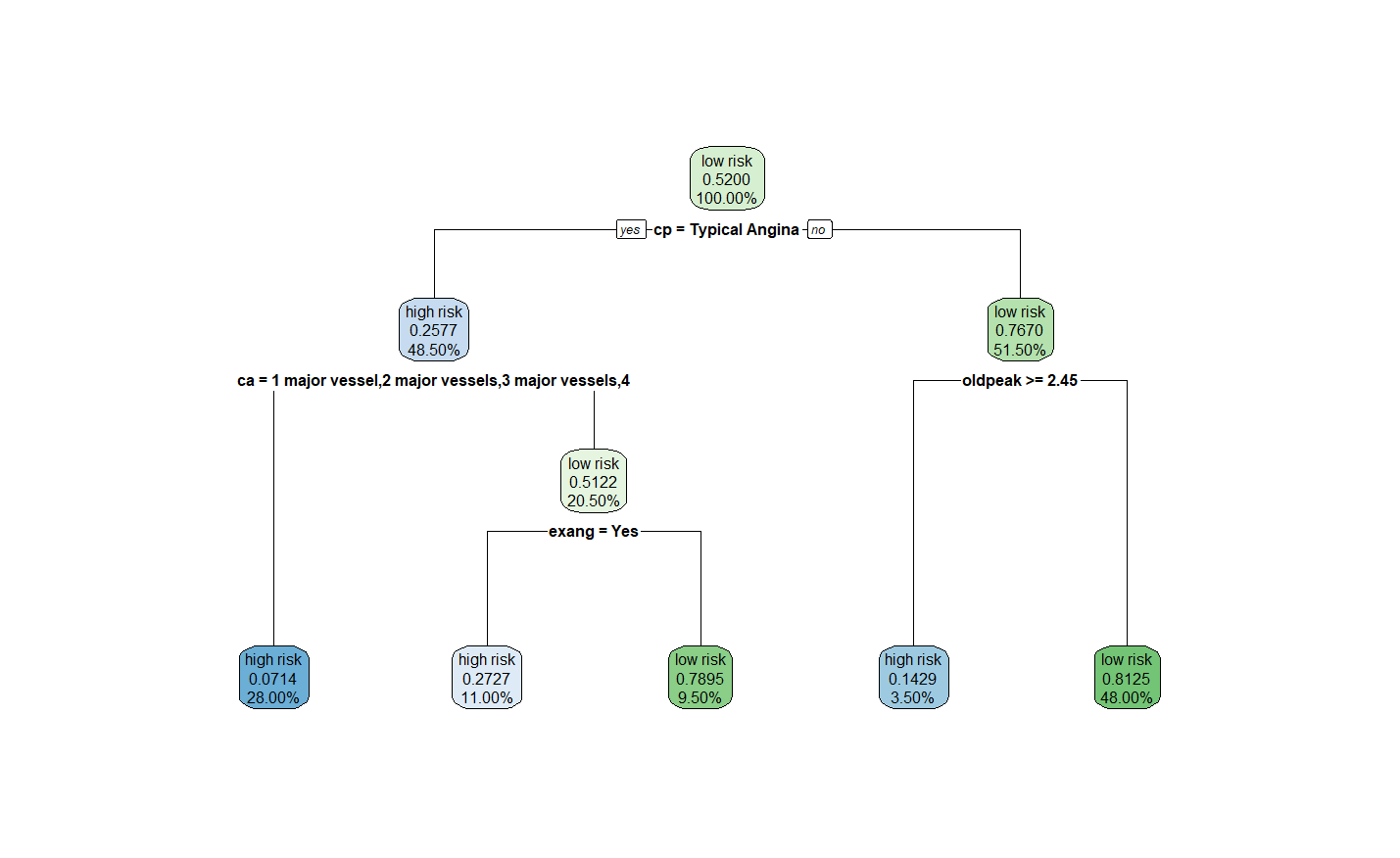
**Figure 8. Decision tree model of demographic variables (age and sex), cholesterol, and blood pressure.**



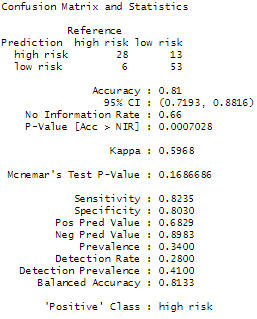
**Image 9. Decision tree model #2 confusion matrix**



**Figure 9. Decision tree model with the high correlation variables.**



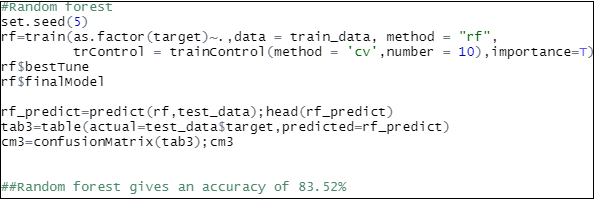
**Image 10. Decision tree model #3 confusion matrix**



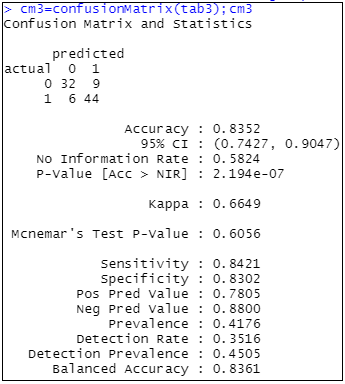
### Random Forest

When running a random forest classification algorithm, we set the seed at five. Our confusion matrix results in prediction accuracy of 83.52%. That is, 76 out of our 91 test data points were predicted correctly.

**Image 11. An image of the code for the Random Forest model.**



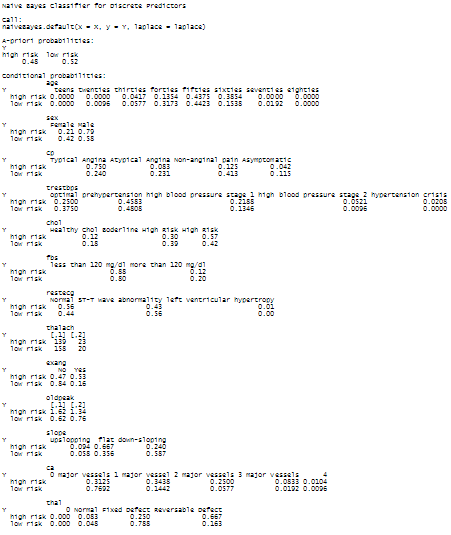
**Image 12. Output for the confusion matrix for the Random Forest model.**



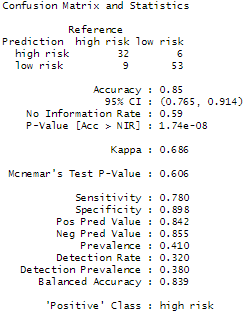
### Naïve Bayes

Similar to the decision tree model, Naive Bayes is a supervised machine learning technique. Nonetheless, this classifier calculates the conditional probability of an event happening (in this case if the patient has a high or low risk of heart disease. In **Image 13,** the conditional probabilities between the target and the rest of the variables in the dataset are calculated. This model achieved an accuracy of 85%.

**Image 13. Naive Bayes Model**



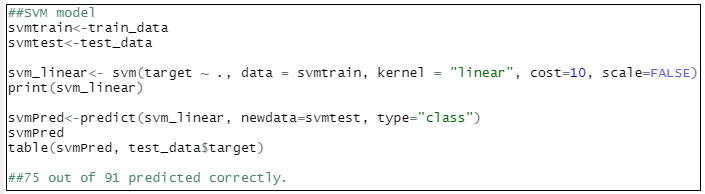
**Image 14. Naive Bayes Model Confusion Matrix**



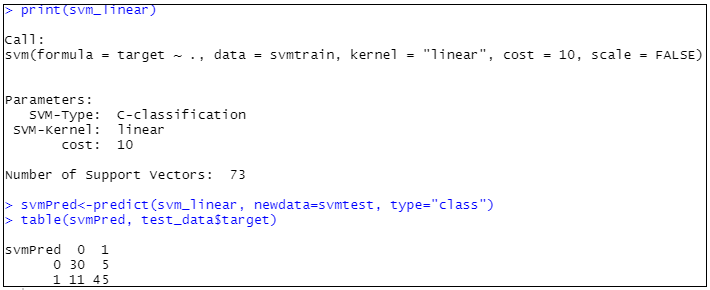
### SVM

Our support vector machine model was run using a linear kernel. The model had an accuracy of 82.42%. That is, it correctly predicted 75 out of 91 test data points.

**Image 15. Image of code for the Support Vector Machine model.**



**Image 16. Output for the Support Vector Machine model.**



# 

# Results

# Association Rule Mining

From the rules seen in **Figure 5**, target patients with a higher chance of developing heart disease are mostly male, are between the ages of forty to sixty, have a resting blood pressure in hypertensive crisis upon hospital admission and have chest pain described as typical angina. These factors found by association rule mining can contribute greatly when examining a patient for heart disease and exhibit some or all of these attributes.

## Clustering / PCA

Clustering is not a predictive method. That said, as discussed above perhaps the most interesting result was the presence of two substantial clusters of people who did not meet the traditional stereotype of people who are high or low risk for heart disease. Additionally, PCA suggested that there is a “general factor of heart disease,” more than merely a bunch of individual factors with predictive power.

## Decision Tree

From **Figure 7**, one could conclude that patients have a higher heart disease risk (30%) if they had a normal or reversible defect status in their thallium test and had typical angina. On the other hand, a patient had lower chances of heart disease if he or she did not have a normal or reversible defect status in their thallium test and a heart rate higher than 148 per minute. Overall, the highest accuracy percent achieved was 81% for the decision tree models. This measure looks promising. The second tree model (**Figure 8**) identified patients in their 60s as the demographic at most risk, and the third tree model (**Figure 9**) identified patients with typical angina and with one or more major vessels colored by fluoroscopy as having a higher risk.

### Random Forest

As shown in **Image 12**, the accuracy for our Random Forest model was 83.52%. It was one of the better predicting models, however, Random Forest in and of itself does not give specific information about the importance of predictive factors within our data set. We can use this model to quickly separate the data into similar groups, but should rely on other models to highlight the more important variables within the data.

**Naïve Bayes**

As shown in **Image 13**, the top variables with high conditional probabilities were: FBS- blood sugar levels (88%, <120mg/d), sex (79%, males), chest pain (75%, typical angina), thallium test status (68%, Reversible defect), slope (67%, flat slope), cholesterol (57%, high), and age (44%, fifties). This model highlighted most of the variables that other models like ARM and decision trees were emphasizing. Compared to the other data mining models performed, the Naive Bayes model scored the highest accuracy (85%).

## 

## SVM

As seen in **Image 16**, our Support Vector Machine model had an accuracy of 82.42%. Similar to our Random Forest model, SVM in itself does not indicate the level of importance of each variable within our data set. While it is helpful to note that there is enough consistency within the data to accurately predict an outcome roughly 83% of the time, other models should be used in support to determine which variables most contribute to our outcome of heart disease development.

# Conclusion

**Table 2.** Overview of model comparisons.

|  |  |  |
| --- | --- | --- |
| Model Comparison | | |
| Model | Accuracy | Significant Attributes |
| ARM | N/A | Sex, Age, Blood Pressure, Heart Rate, Chest Pain, Major Vessels Colored |
| Clustering / PCA | N/A | Main axes: overall heart disease risk, diabetes risk, |
| Decision Trees | 80% | Thallium Test Status, Chest Pain |
| Naive Bayes | 85% | Blood Sugar Levels, Sex, Chest Pain |
| Random Forest | 83.52% | N/A |
| SVM | 82.42% | N/A |

Overall, Naive Bayes was the most successful data mining method employed with an accuracy of 85%. Both the Random Forest model and the Support Vector Machine model can help to support the notion that there is enough consistency within our data to accurately predict our outcome variable roughly 83-84% of the time. In addition to Naive Bayes, the decision trees also indicated some of the important variables within the data set. The Naive Bayes model determined that the most significant attributes within the data set are blood sugar levels, sex, and presence of chest pain. The decision trees determined the important variables to be the thallium test status as well as the presence of chest pain.

The existence of chest pain was highlighted as a significant variable in at least two of the models (Naive Bayes and decision trees). This is a factor which the American Heart Association lists as the number one symptom of heart attack. In fact, pain is both the number one and number two symptoms of heart attack (number two being pain in the arm, back, neck, jaw, or stomach). It is reassuring to see that our models could accurately predict one of the most if not the number one most important factor in the prediction of heart disease.

# Bibliography

ARM: Hahsler, M. (n.d.). Introduction to arules – A computational environment for mining association rules and frequent itemsets. Retrieved August 31, 2020, from https://cran.r-project.org/web/packages/arules/vignettes/arules.pdf

Cholesterol levels: U. (2013). Different types of cholesterol, Retrieved August 31, 2020, from http://www.statinusage.com/Pages/cholesterol-overview.aspx

Dataset: Naresh (2020). Health care: Data Set on Heart Attack Possibility (Version 1) [CSV file]. Retrieved September 10, 2020 from https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility.

Blood Pressure levels: American Heart Association. (n.d.). Understanding Blood Pressure Readings. Retrieved August 31, 2020, from https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Heart Attack Symptoms in Women (July 31, 2015). Retrieved September 10, 2020, from https://www.heart.org/en/health-topics/heart-attack/warning-signs-of-a-heart-attack/heart-attack-symptoms-in-women

GitHub Repository: https://github.com/larregui/IST-707-Project.git