Project Title: Heart Disease Classification and Exploration

# Technical Analysis:

## Intro to Dataset and Preprocessing:

For our project we decided to work on a dataset regarding heart disease diagnosis prediction The dataset we worked with is the Cleveland Heart Disease dataset taken from the UCI repository. This data set consists of 303 individuals, 13 predictor variables and 1 binary classification variable (target) for a positive or negative for Heart Disease diagnosis. Our goal is to analyze which factors are the most effective at predicting heart disease risk as well as what latent factors we can discover from within the data set that may help us to better understand the relationships between the predictors.

Our analysis consisted of three main portions: A regularized regression to predict diagnosis, a factor analysis to understand our predictors and a creating a second regularized regression to compare to using the factor scores taken from our factor analysis.

Prior to running our analyses we had to do some pre-processing. After some domain research we determined our predictors consisted of the following types:

* Binary (4): sex, fbs, exang, target
  + Target is the diagnosis variable (1=Positive Diagnosis, 0 = No diagnosis)
* Ordinal (5): cp, restecg, slope, ca, thal
* Numeric (5): age, trestbps, chol, thalach, oldpeak

Our data has three types of predictors: Binary, Ordinal, and Numeric. The Ordinal variables have either 3 or 4 categories. Our dependent variable “target” is also a binary.(1=Positive Heart Disease Diagnosis, 0 = Negative diagnosis). The meaning of each variable can be found in the appendix [(A)](#hq3ihaw8cunt).

Referring to our variable distributions ([A1](#tebx1cmjlfom)), we can see all variables follow a normal distribution or have balanced categorical distributions, with the exception of chol ([A2](#giqiol6dz9aq)), oldpeak ([A3](#zcsy8kl8t7cv)), and trestbps [(A4](#fw9oz9j1wauf)). We Removed outliers by the following criteria: oldpeak > 4, trestbps > 180 and chol > 450. This only removed only 8 points in total.

## 

## 1.Relaxed Lasso Logistic Regression

For our first analysis, I decided to run a Logistic Lasso Regression with the main goal of creating a predictive model to determine whether future patients held a diagnosis of heart disease. Furthermore, I had hoped to create a model with numbers that would allow a doctor to create an initial diagnosis, so that the expense of future testing could be reduced for the patient. However, normal Lasso regression has an issue of introducing Bias to the models that are created. This is due to normal Lasso regression applying L1 regularization, which forces some coefficients to be exactly zero, effectively selecting a subset of features. However, this introduces bias in the coefficients selected because Lasso shrinks even the nonzero coefficients toward zero. Therefore, to combat this bias introduced, I decided to use Relaxed Lasso regression. This uses Lasso to find the variables important for the model, but then re-estimates the variables to reduce this over-shrinkage. Therefore, by introducing Relaxation, the model is less biased, while increasing predictive power.

After having the method planned, I then began by creating two different sets of data, these being a training and testing set. As our model is to be used to find its predictive ability, having a testing set is necessary for complete analysis. For splitting the dataset into two different sets, I used a 70/30 split. Although the dataset could be defined as large and small, compared to some other massive datasets, I felt that a 70/30 split allowed for a middle ground to be held. This meant that my training set had 210 patients and my testing set had 93 patients. After splitting the data, I once more had to create two new sets of data for both the training and testing sets. However, this was for the regression technique as Lasso regression requires the independent variables and dependent variable separate. Therefore, the diagnosis variable was separated from the rest.

After doing these steps, I was then able to run Relaxed Lasso Regression and plotted this technique ([A5](#phsu11e4fmfu)). As one can see, the plot indicates that gamma=.25, which is represented by the green line, achieves the lowest binomial deviance. This suggests that this gamma provides the best balance between predictive accuracy and model parsimony. This allows me to have only the most important variables to increase explainability while achieving a strong predictive performance. Finally, the plot supports the use of relaxed lasso over standard lasso, as it reduces the shrinkage needed.

After selecting gamma=.25 as the optimal relaxation parameter, I then choose lambda.1se to determine the final set of coefficients and variables ([A6](#xxwq5bcss3af)). The variables selected in this process were the type of chest pain, exang (chest pain which is trigged from exercise), CA (number of major vessels affected by the extent of coronary blockages), and Thal (representing the severity of thalassemia, a blood disorder). After this model was created, I then looked to focus on the accuracy at which this model predicts diagnosis ([A7](#yriwarmv0wvy)). This produced a model accuracy of approximately 89% ([A8](#yriwarmv0wvy)). When looking at the accuracy table, one particular number is extremely important. This number is defined as a False Negative which is when the actual diagnosis is True but the model predicts that the diagnosis is False. This is extremely dangerous is models that deal with any disease that can be fatal. As our dataset relates to the heart, the most important organ in your body, having a false negative could mean fatal results if no further testing is done. Therefore, our model above held an approximate 4.5% rate of false negatives. This is good for normal models; however, this could be seen as too high for a model that is testing for something as important as the heart.

However, another aspect that needs to be looked at during this process of Relaxed Lasso Logistic Regression is the way that samples are created. Since the process of finding random samples is different for every time, creating different sets of samples meant that different models were created from the dataset. Another model that was produced can be seen below ([A9](#l9mvovocme78)). Unlike the previous model, this model has the variables thalach (the maximum heart rate achieved) and oldpeak (oxygen deprivation during states of exercise). The variables exang (chest pain which is trigged from exercise), CA (number of major vessels affected by the extent of coronary blockages), and Thal (representing the severity of thalassemia, a blood disorder) are all still present but this new model dropped the type of chest pain. This new model has extremely similar results as well when it comes to prediction accuracy ([A10](#z5wbdwgk9wgk) & [A11](#9rr37ni2jfub)).

Although what samples are created may be different every time you run the piece of code, the results after many models are created are all connected. The variables that are most present in the models after many different attempts is type of Chest Pain, Thalach, Exang, Oldpeak, CA, and Thal; where the meaning behind these variables can be seen above.

## 2.Factor Analysis:

For our second analysis I decided to run a Factor Analysis (FA) with the main goal of attempting to determine latent factors that can help us to create a comparable LASSO logistic regression model to classify Heart Disease. In addition, I hoped this would help us to better understand the relationships that our predictors are telling us. Due to the combination of Numeric, Binary and Ordinal Variables I determined using a mixed correlation ([A12](#n7ovmcdxy2ry)) of Polyserial, Polychoric and Pearson was the proper correlation to use for our scaled Factor Analysis, since typical measures like Pearson, Spearman’s Rank, and Kendall’s Tau would not work. We can gather a few initial insights from this plot, For example, the strongest correlation in our analysis is between sex (coded as 1 for male) and thal (representing the severity of thalassemia, a blood disorder). This suggests that males may be significantly more likely to have thalassemia than females*.* Furthermore, we see that the maximum heart rate achieved (represented by thalach) has a negative correlation with almost everything else, meaning that it may become its own factor. In addition some of the other stronger correlations we have are between: slope and oldpeak, cp and exang as well as between age and ca. This led me to expect that these pairs may be placed in similar factors. We do lack any sort of visible clusters which makes me worry about the validity of our Factor Analysis.

I ran the first few FAs, with both the minimum residual methods (EFA) as well as the Maximum Likelihood method (CFA) and played around with the cutoff and number of factors. After comparing those preliminary scaled FA’s to a parallel analysis ([A13](#ogi38x229bvs)), I determined that 5 factors is the optimal number of components based on the criteria of variance = 1. While the knee comes after the second factor, we need to allow for more factors to capture enough variance to ensure our FA is suitable.

I ran the next sequence of FA’s all with 5 factors and a cutoff of 0.5. The stricter cutoff of 0.5 was decided upon because of the overall low individual correlations and I had the goal of reducing cross loadings and making interpretable, parsimonious factors. A Common Factor Analysis (CFA) using no rotation ([A14](#txmnuwyzjkt)) resulted in factors having overlapping variables and the total variance was only 0.527. Overlapping variables are a problem because if variables load on multiple factors, it becomes more difficult to interpret the meaning of the factors clearly. Additionally, if the total variance explained by the factors is low, it suggests that the extracted factors do not adequately represent the underlying data. After this I ran a second Common Factor Analysis using a VARIMAX rotation ([A15](#jwkds6odirkz)), with the hopes of forcing the variables into distinct factors. This worked as the resulting factors had distinct variables, but certain predictors were lacking representation in factors. The total variance was still only 0.527.

Following these spotty results I decided to examine the suitability of my data for Factor Analysis before I moved any further. After running Bartlett’s Test of Sphericity ([A16](#hyfwc7wus8z6)) I got a p-value < 2\*10^-16 meaning that my correlation matrix is significantly different from an identity matrix (where all variables are uncorrelated), meaning there are strong correlations between variables, making factor analysis appropriate. I then ran a KMO measure of sample adequacy ([A16](#hyfwc7wus8z6)) on the entire dataset and got a score of 0.646, which was mediocre at best. I then examined the individual KMO scores of each predictor to see if there were any that I may consider removing. sex (0.424), chol (0.478), restecg(0.500), trestbps(0.517), fbs (0.550) were all extremely weak, meaning each variable does not contribute much to the FA. I decided to remove them and re-run these tests. After removing those 5 predictors, our overall KMO increased to 0.769 and all of the individual KMO measures were above 0.7 ([A16](#hyfwc7wus8z6)), indicating much greater sample adequacy both overall as well as among our remaining individual predictors.

I then re-ran my Common Factor Analysis with a Varimax rotation and those 5 predictors removed ([A17](#vuvyeeluci5)). This produced all distinct factors, all remaining variables are represented in factors, and our total variance has increased to 0.631, which is not incredibly impressive, but is a large improvement from prior to the variable selection.

Our most significant factor (ML5) I named “Heart Health During Exercise”. This represents cardiac stress response (slope) and oxygen deprivation (oldpeak) particularly during states of exercise. These are both critical indicators of heart disease, and a higher score for this factor could indicate that.

Our second most significant factor (ML3) I named “Chest Pain Severity”. This represents symptoms of angina, which is a type of chest pain caused by reduced blood flow to the heart. This component captures how severely chest pain manifests (cp) and whether it is triggered by exercise (exang). This demonstrates that angina can also be a sign of heart disease so a higher score here could indicate higher chance of heart disease diagnosis.

Our third most significant factor (ML2) I named “Anatomical Heart Structure & Age”. ML2 has less of a connection between its predictors, but still captures variables that impact a higher chance of heart disease. The number of major vessels affected by the extent of coronary blockages (ca) and years lived (age) end up in the same factor. This tells us that as people get older their risk for coronary blockages increases and that the combination of both may lead to greater risk of heart disease. While the predictive power is lower, this factor is a bit less relevant to heart disease diagnosis, but as a latent factor this still allows us to understand another relationship among our predictors.

Our next factor ML4 consists of a single variable (thalach), which represents the maximum heart rate achieved or “Heart Rate Efficiency”. Its interpretable significance comes from its negative correlation, meaning that people with a lower max heart rate tend to have worse cardiovascular health. So on the contrary, a high score for ML4 may indicate a healthier patient.

Our final factor also consists of a single variable (thal) which represents the presence of thalassemia, an inherited blood disorder that affects the body's ability to produce hemoglobin, the protein in red blood cells that carries oxygen, and its potential effects on the cardiovascular system. A higher score for this factor indicates worse thalassemia which could indicate a higher chance at heart disease and a lower score indicates the opposite.

If we revisit the original mixed correlation matrix we can see the inner square that has most of the more highly correlated variables, but not many clumps. We did have to remove multiple predictors and as a result we never had factors with more than only two variables in them. While we were able to see interesting latent factors, I am still unsure about our predictive power that the scores of this factor analysis will have.

## 

## 3.Regularized Regression - From Factor Analysis

I then used the factor scores for the factors with multiple variables to create a LASSO regression model ([A18](#u10e83cj289s)). Due to computational issues, I included the variables that were their own factors as predictors in their variable state. We can see that ML3 “Chest Pain Severity”, has a coefficient estimate of 1.613, making it our strongest coefficient and thus will play the most important role in classifying heart disease. Since we are doing logistic regression, transforming we get (e^1.613 = 5.02), which means that for a 1-unit increase in ml3, the odds of a positive heart disease diagnosis occurring increase by a factor of 5.02. Additionally Thal (0.910) ended up being our second strongest predictor, meaning that the prevalence of Thalassemia will also play a strong role in predicting a heart disease diagnosis. (e^0.910 = 2.48), which means that for a 1-unit increase in thal, the odds of a positive heart disease diagnosis occurring increase by a factor of 2.48. These coefficients are a bit different than what we could have expected based on the importance given from our factor variances, so it leads me to continue to be skeptical of the power of this predictive model.

Using my newly created LASSO model I computed predictions on the entire dataset ([A19](#huljni2ub91r)). While I got an accuracy of about 84%, We have double the amount of false negatives than we do false positives. This is incredibly worrisome as diagnosing someone who actually has heart disease with a negative diagnosis is extremely problematic as they may not receive the proper treatment they need. This could be due to the lack of variance captured by our data as we did remove multiple predictors that may have been needed to improve those missed calculations. In addition the high variance may be due to the fact that we are computing predictions on the same dataset that our model was built on, so the model would have an inherent tendency to overfit.

While we may have been getting interesting latent factors and a comparable predictive accuracy on the LASSO regression built on the Factor Scores, our results for Factor Score LASSO are much worse than the traditional LASSO regression because of the result of extremely high false negatives. The interpretability of the latent factors helps us to understand how we could further evaluate and begin to potentially detect heart disease early on, but as predictors for a diagnosis, they do not yield strong results.

## 4.Conclusions

While these two or three statistical models provide powerful models for identifying key predictors in heart disease, their effectiveness is inherently limited by their accuracy. When looking at the Lasso Approach we can see an average of 88% accuracy when choosing a gamma of .25 and selecting lambda.1se. Similarly when creating a predictive model from the factor analysis we saw an accuracy of about 84% with a concern more on the number of false negatives. These methods are valuable in our medical research, guiding for future decision-making and identifying at-risk patients.

However, when it comes to human lives, these models with approximately 85-90% accuracy still leave too much room for dangerous misclassifications. With these models, a missed diagnosis should occur in 1 out of every 10 patients which is extremely dangerous. Therefore, with these models in mind, there is much room for further progress and better models to be made. Our goal is not simply to create effective models when dealing with human lives. Our goal is to hopefully build near-perfect models that will help with the future of saving lives. Therefore, with all the models built, testing multiple times, even though cost may be high, can be concluded when dealing with a person’s life.

# Non-Technical Analysis (1-2 pages):

For our project, Liam and I decided to examine certain traits about medical patients with the goal of being able to predict their medical diagnosis as well as examine the relationships between certain medical traits. We used two main routes of analysis, one that used the importance of the individual traits themselves and one that examined how the traits could be combined in order to improve our understanding of how they interact with each other:

To begin, we first began by analyzing the importance of the individual traits within the dataset; however, there are many different paths for which this could occur. Therefore, we choose to create a predictive model in the hopes to find a model with high accuracy for future patients. Thus allowing us to reduce the cost for patients.

The model we found had some extremely interesting results. To begin with, the models produced on average found an accuracy of 88%. One Specific model even found an approximate 90% accuracy ([A10](#z5wbdwgk9wgk)). Furthermore, a fear of ours was that of false negatives. A false negative occurs when the predictive model places the diagnosis as False, when the true diagnosis is True during the testing phase. The models found an approximate 4.5% occurrence rate of false negatives. This is not amazing for diseases that could be fatal.

However, the key individual traits that can be found from these models are type of Chest Pain, Thalach (the maximum heart rate achieved), Exang (chest pain which is triggered from exercise), Oldpeak (oxygen deprivation during states of exercise), CA (number of major vessels affected by the extent of coronary blockages), and Thal (representing the severity of thalassemia, a blood disorder)

From analyzing the different combinations of medical traits we examined a few intriguing relationships between certain traits. We were able to create combinations of predictors that we could use to evaluate certain participants in order to determine their likelihood of obtaining a heart disease diagnosis.

For example, certain traits focused on analyzing how heart health changed during exercise, we found that the rate of cardiac stress response and oxygen deprivation during exercise were related with one another and together they could be used as one of a predictor in determining a heart disease diagnosis.

Another commonality we found was the impact of chest pain on heart disease diagnosis.We examined how severely chest pain manifests and whether it is triggered by exercise. This led us to further understand that by analyzing metrics related to exercise, we could get a better idea about a patient’s overall heart health. We also learned that increased chest pain may result in worse performance during exercise.

Another one of the most interesting takeaways we found was the impact of age on a certain heart condition: the number of coronary blockages. We found that increased age likely means that a patient would have a greater number of coronary blockages in their heart and the combination of these two can give us insight into a patient's risk of heart disease.

Two other takeaways we made were that Heart rate efficiency alone is a good measure to use to evaluate a patient, as those with a lower max heart rate tend to have worse cardiovascular health and vice versa. In addition, the severity of other related heart conditions such as thalassemia can impact heart disease likelihood as well.

Using these newly created combinations of the variables, we attempted to examine how accurately we could predict heart disease compared to our first method of analysis. While we got a prediction accuracy of about 84% across all patients, we have a severe problem with the ability to correctly diagnose those who actually have heart disease. This is incredibly worrisome as mis-diagnosing someone who actually has heart disease with a negative diagnosis is extremely problematic as they may not receive the proper treatment they need. This leads us to lean more favorably to Liam’s predictive measures rather than mine.

However, the analysis I ran was still beneficial because it helps us to understand how we could further evaluate and begin to potentially detect heart disease early on by the combination of certain patient traits. It helps us to further our understanding of all of the different traits that impact heart disease and how they interact with each other.

# 

When combining the two analysis methods, we can see positive results. Whether at the individual level or the combination level, many positives can be taken away; as we found many characteristics that are related to heart disease. That being said though there can be some conclusions made, but a true model was not found. When dealing with human life, sometimes numbers and models can be at fault, in particular with a false negative. These models at best will still have one prediction wrong for every 10 patients. Therefore, further testing should always be done to make sure that the life of the patient is safe. That being said, we now have ways of better predicting the diagnosis, allowing for fewer tests to be necessary.

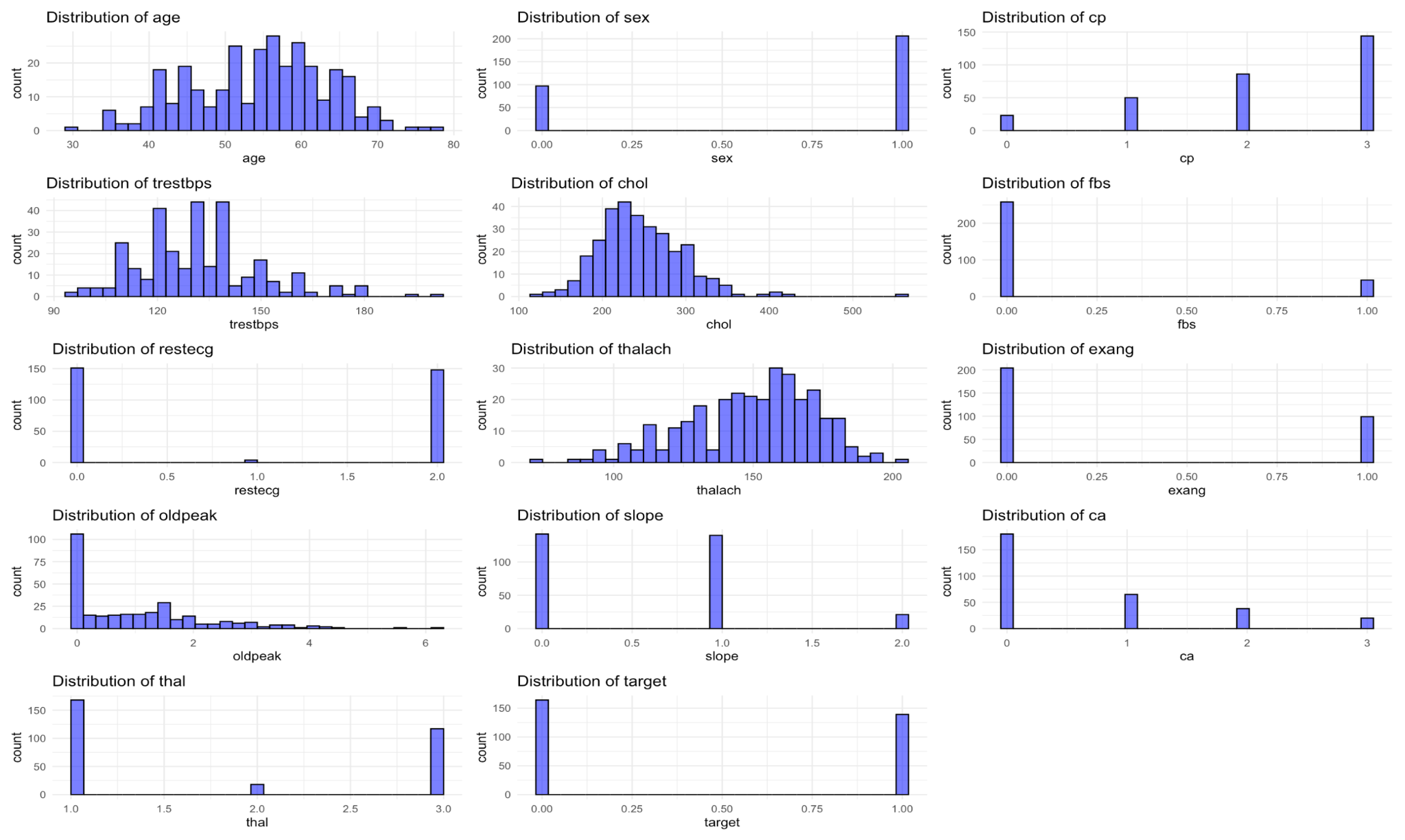
# Appendix:

A

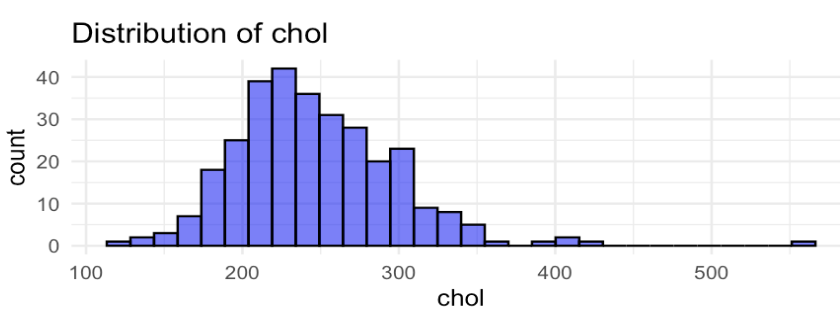
# Variables Explanations:

* **age –** (Numeric)**:** Patients Age in years
* **sex –** (Binary) **:** Gender (Male : 1; Female : 0)
* **cp –** (Nominal) **:** Type of chest pain experienced by patient. This term is categorized into 4 categories.
  + 0 typical angina,
  + 1 atypical angina,
  + 2 non- anginal pain,
  + 3 asymptomatic
* **trestbps –** (Numeric) **:** patient's level of blood pressure at resting mode in **mm/HG**
* **chol –** (Numeric)**:** Serum cholesterol in **mg/dl**
* **fbs –** (Binary)**:** Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false
* **restecg –** (Nominal)**:** Result of electrocardiogram while at rest are represented in 3 distinct values
  + 0 : Normal
  + 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  + 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
* **thalach –** (Numeric)**:** Maximum heart rate achieved
* **exang** – (Binary) **:** Angina induced by exercise 0 depicting NO 1 depicting Yes
* **oldpeak –** (Numeric)**:** Exercise induced ST-depression in relative with the state of rest
  + 0 is baseline good, higher means the heart isn’t getting enough oxygen during exercise
* **slope –** (Nominal)**:** ST segment measured in terms of slope during peak exercise
  + 0: up sloping;
  + 1: flat;
  + 2: down sloping
* **ca –**  (Nominal)**:** The number of major vessels (0–3)
* **thal –** (Nominal)**:** A blood disorder called thalassemia
  + 0: NULL
  + 1: normal blood flow
  + 2: fixed defect (no blood flow in some part of the heart)
  + 3: reversible defect (a blood flow is observed but it is not normal)
* **target** – (Binary)**:** It is the target variable which we have to predict 1 means patient is suffering from heart disease and 0 means patient is normal.

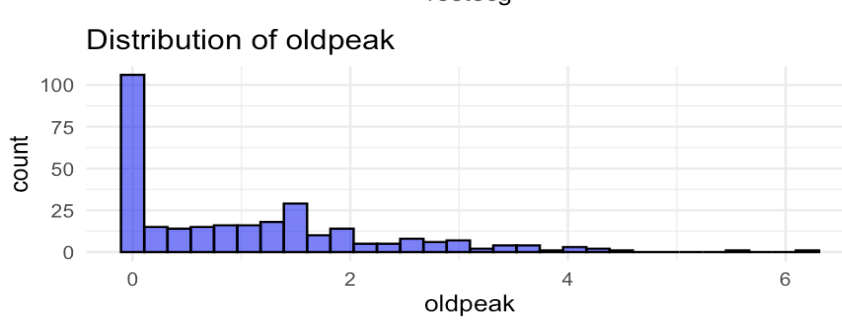
A1



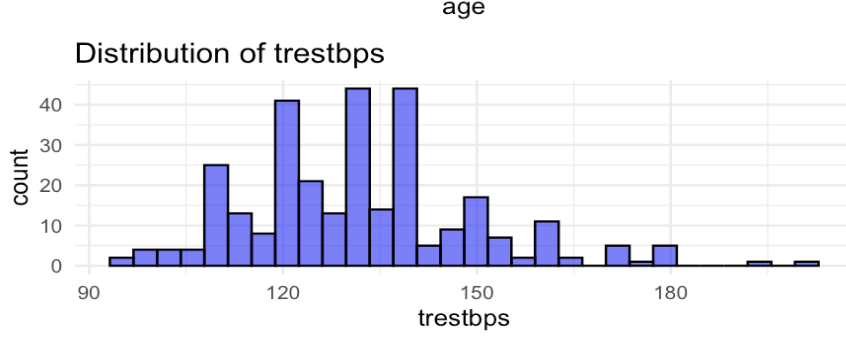
A2



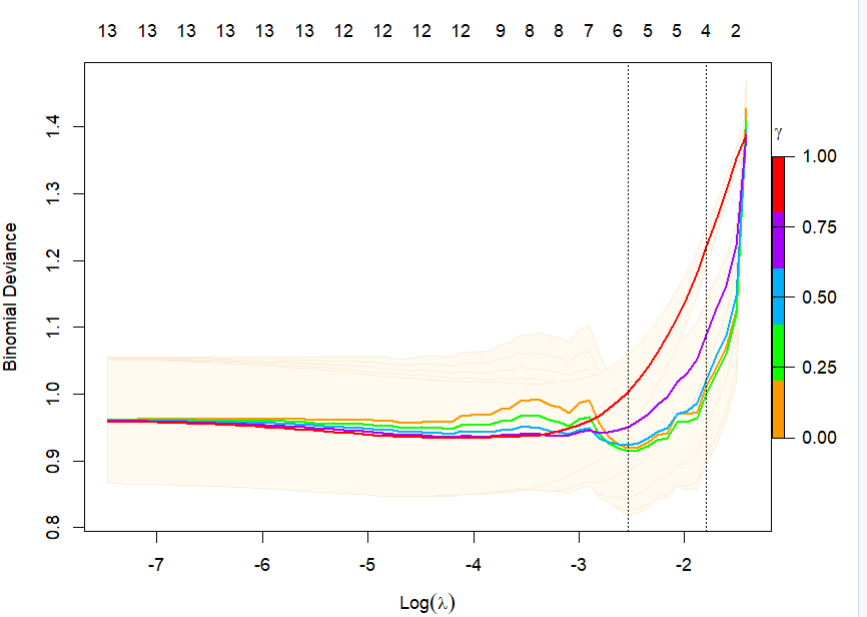
A3



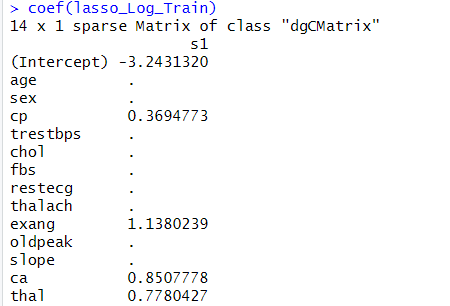
A4



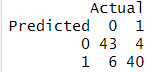
A5



A6



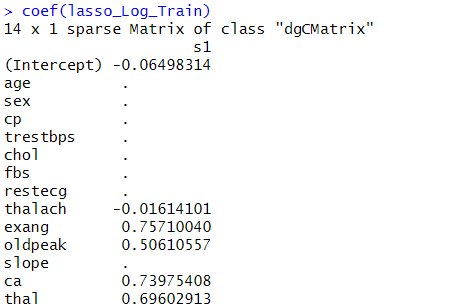
A7



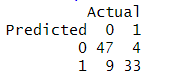
A8



A9



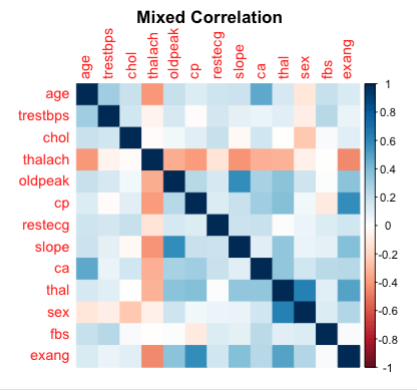
A10



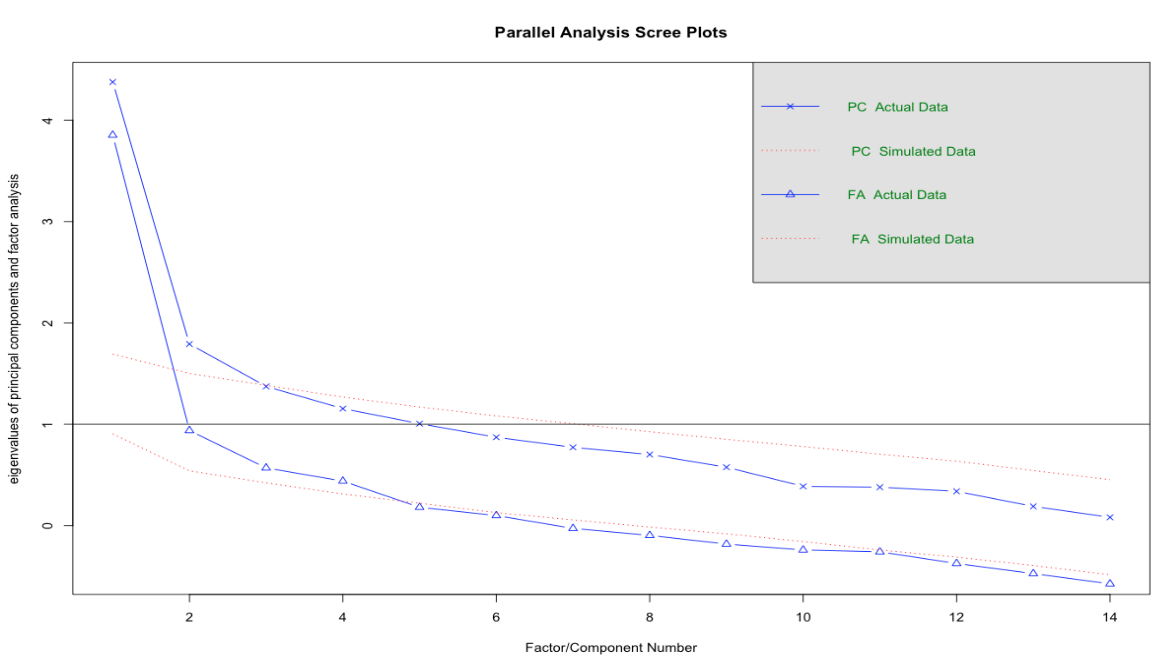
A11



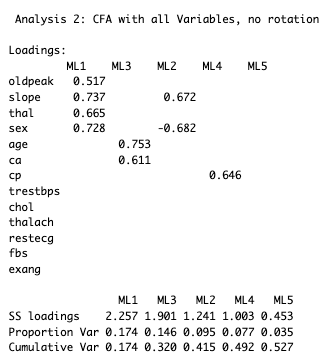
A12



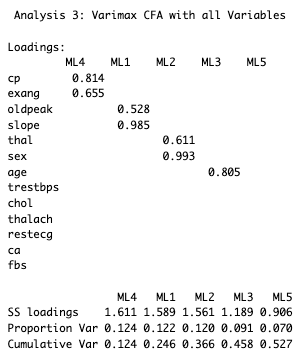
A13



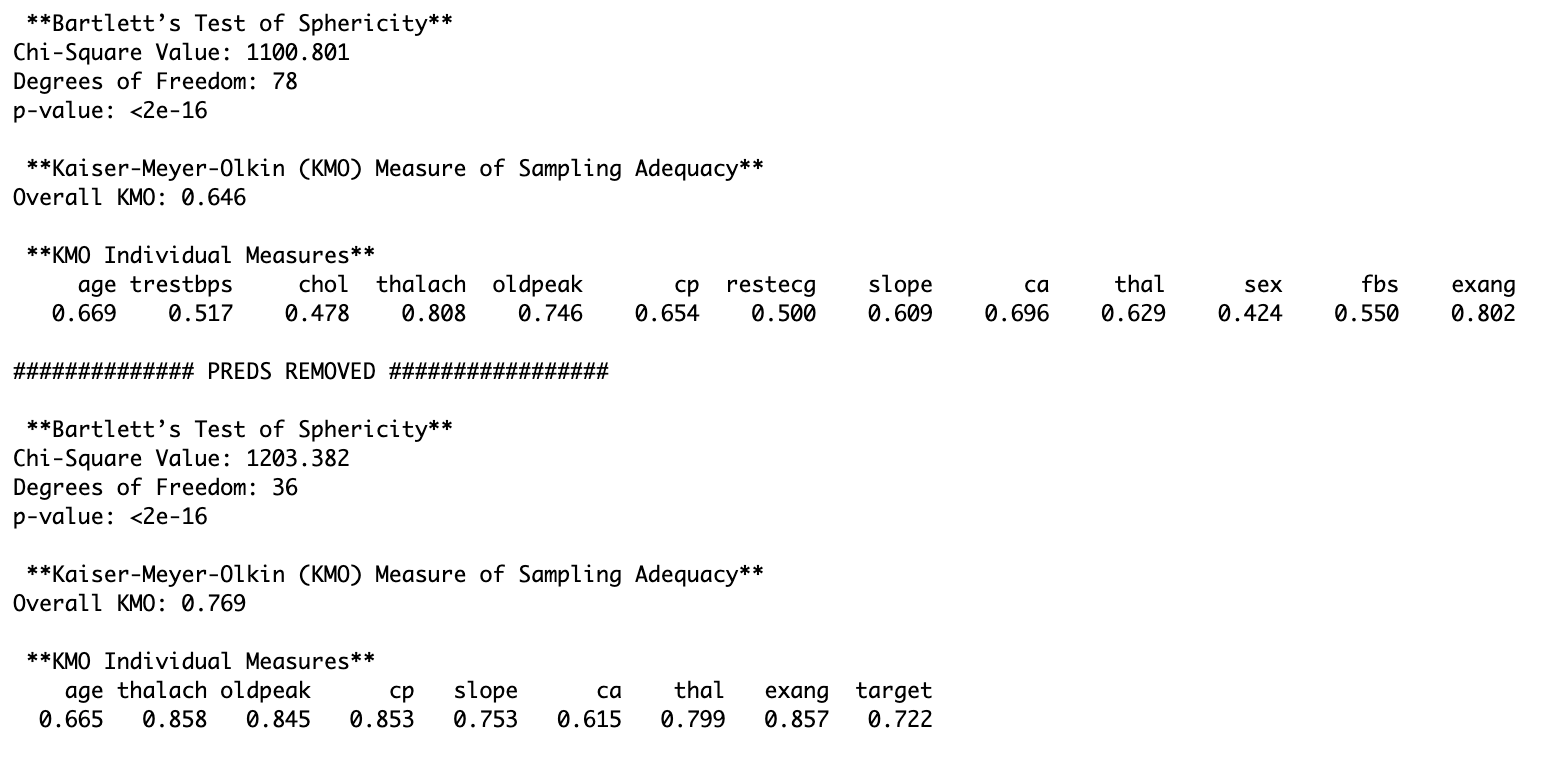
A14



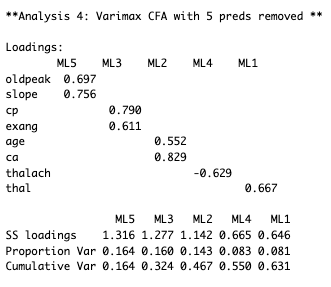
A15



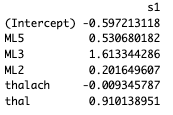
A16



A17



A18



A19

