**Shrinkage Methods and Tree-Based Methods**

**1. INTRODUCTION**

1. Questions of Interest:

For this portion of our project, our group will be investigating two questions: a regression question and a classification question.

1. **Regression:** What is the relationship between a patient’s level of the CPK enzyme (creatinine phosphokinase) and their personal demographics and health history?
2. **Classification:** Are there different health characteristics for patients who die from heart failure than for those who do not die from heart failure?

We will be investigating several variables related to the response variables of “creatinine\_phosphokinase” and “DEATH\_EVENT,” exploring the ways in which they relate to the level of CPK enzyme and the ways in which they differ for those patients recorded as dying, and those recorded as not dying.

1. Question Importance:

Firstly, for our regression question, high levels of the CPK enzyme can indicate that there are internal damages to a patient’s brain or muscle tissue (Chen). This enzyme is particularly important to understand when it comes to heart disease, as abnormal levels of CPK are often associated with conditions such as myocarditis, cardiomyopathy, and heart attacks (Chen). Therefore, in addition to understanding what high levels of this enzyme mean for a patient’s health, it is equally as important to investigate behavioral and health characteristics that can lead to its increase in the body. Thus, we believe it is worth analyzing the ways in which the variables in this dataset are associated with heightened CPK levels. Overall, if our analysis shows that there are strong correlations between the demographic and health characteristics of a patient and the level of CPK in the blood, we can potentially glean insights into ways to prevent the enzyme from spiking as a result of damaged muscle tissue.

Secondly, for our classification question, heart failure is a common cause of death, with almost 380,000 mentions on death certificates in 2018 in the United States alone (CDC). Heart failure also has a large financial impact, costing around $30.7 billion in 2012 in medicine and healthcare costs (CDC). Further research on risk factors can lead to earlier treatment, which would decrease both the death rate and the financial cost of heart failure. An analysis that pinpoints the main predictors of heart failure will help physicians estimate the direness of a patient’s condition. This will allow them to allocate resources to the most serious cases and save lives of people with high risk factors. This knowledge could also help public servants pursue beneficial health policy. If a predictor is especially correlated with death by heart failure, policymakers can redirect funding to research the root cause of that predictor and how to alleviate the issue.

1. Dataset

In order to explore the aforementioned question of interest, our report will use a dataset called “Heart Failure Prediction.” We found this free dataset on the website, Kaggle. The user obtained the original source material from BioMed Central (also known as BMC), which is owned by the medical research publisher, Springer Nature. The dataset is made up of information from the medical records of nearly 300 heart failure patients at the Faisalabad Institute of Cardiology in Pakistan, all of which was collected during the year 2015. The dataset includes information about the lifestyle habits and clinical health of each patient, describing the potential risk factors and predictors of death from heart failure. For this analysis, in addition to the binary, categorical response variable, we will be using 5 categorical variables and 6 quantitative variables from the dataset.

1. Variable Description

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| DEATH\_EVENT | Whether or not the patient died from heart failure | Categorical, binary  *- Levels: Death, No Death* |
| Age | Age of the patient in years | Quantitative, continuous |
| Anaemia | Patient’s anemia status - whether or not the patient has anemia (decreased hemoglobin) | Categorical, binary  *- levels: Anemic, Non-Anemic* |
| Creatinine\_phosphokinase\*\* | Level of the CPK enzyme in the blood (mcg/L) | Quantitative, continuous |
| Diabetes | Patient’s diabetes status - whether or not the patient has diabetes | Categorical, binary  *- levels: Diabetic, Non-Diabetic* |
| Ejection\_Fraction | Percentage of blood leaving the heart at each contraction | Quantitative, continuous |
| High\_blood\_pressure | Patient’s hypertension status - whether or not the patient has hypertension | Categorical, binary  *- levels: Hypertensive, Non-Hypertensive* |
| Platelets | Level of platelets in the blood (kiloplatelets/mL) | Quantitative, continuous |
| Serum\_creatinine | Level of serum creatinine in the blood (mg/dL) | Quantitative, continuous |
| Serum\_sodium | Level of serum sodium in the blood (mEq/L) | Quantitative, continuous |
| Sex | Patient’s gender - whether the patient is male or female | Categorical, binary  *- levels: Male, Female* |
| Smoking | Patient’s smoking status - whether or not the patient smokes | Categorical, binary  *- levels: Smoker, Non-Smoker* |

\*\* indicates the response variable

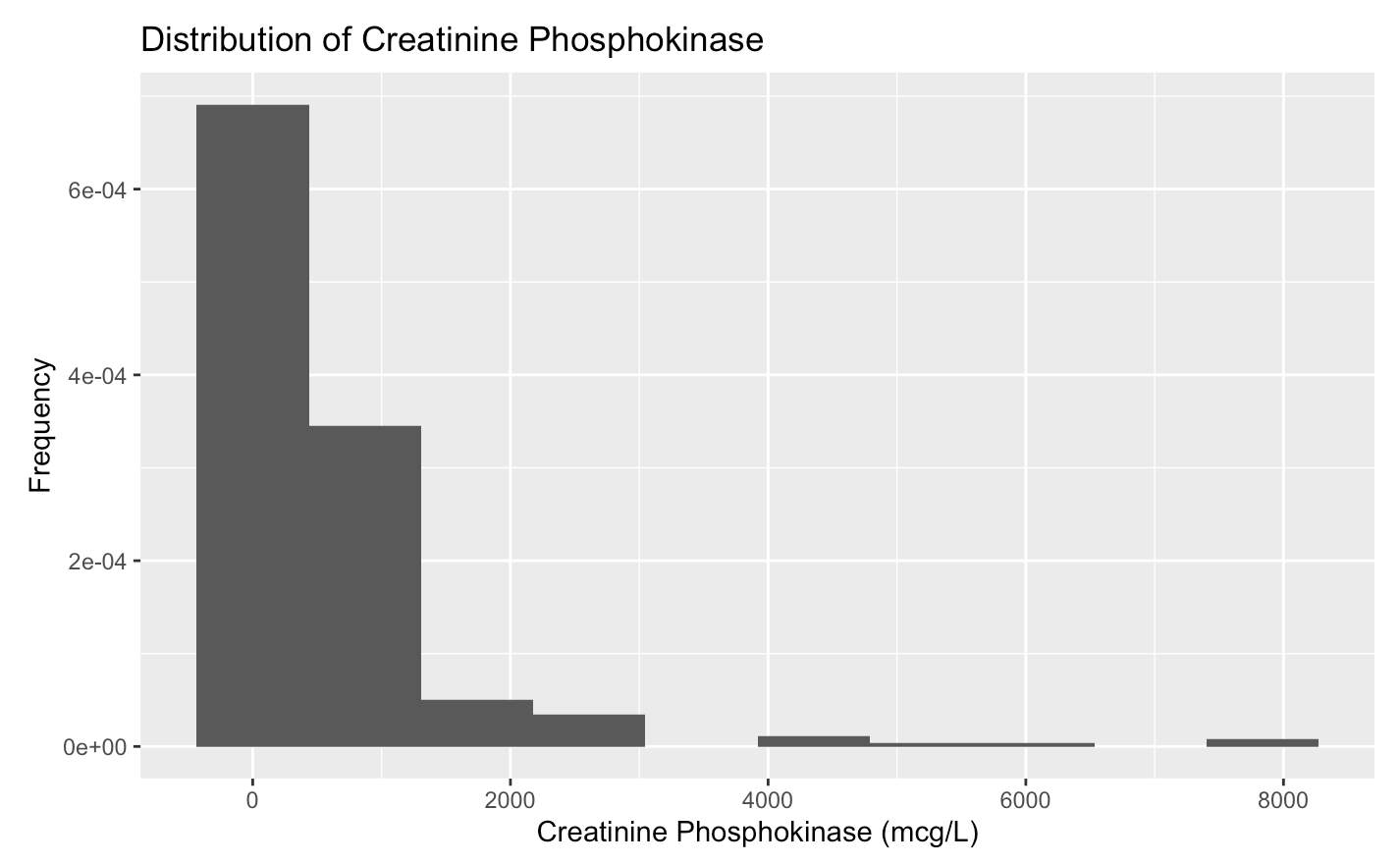
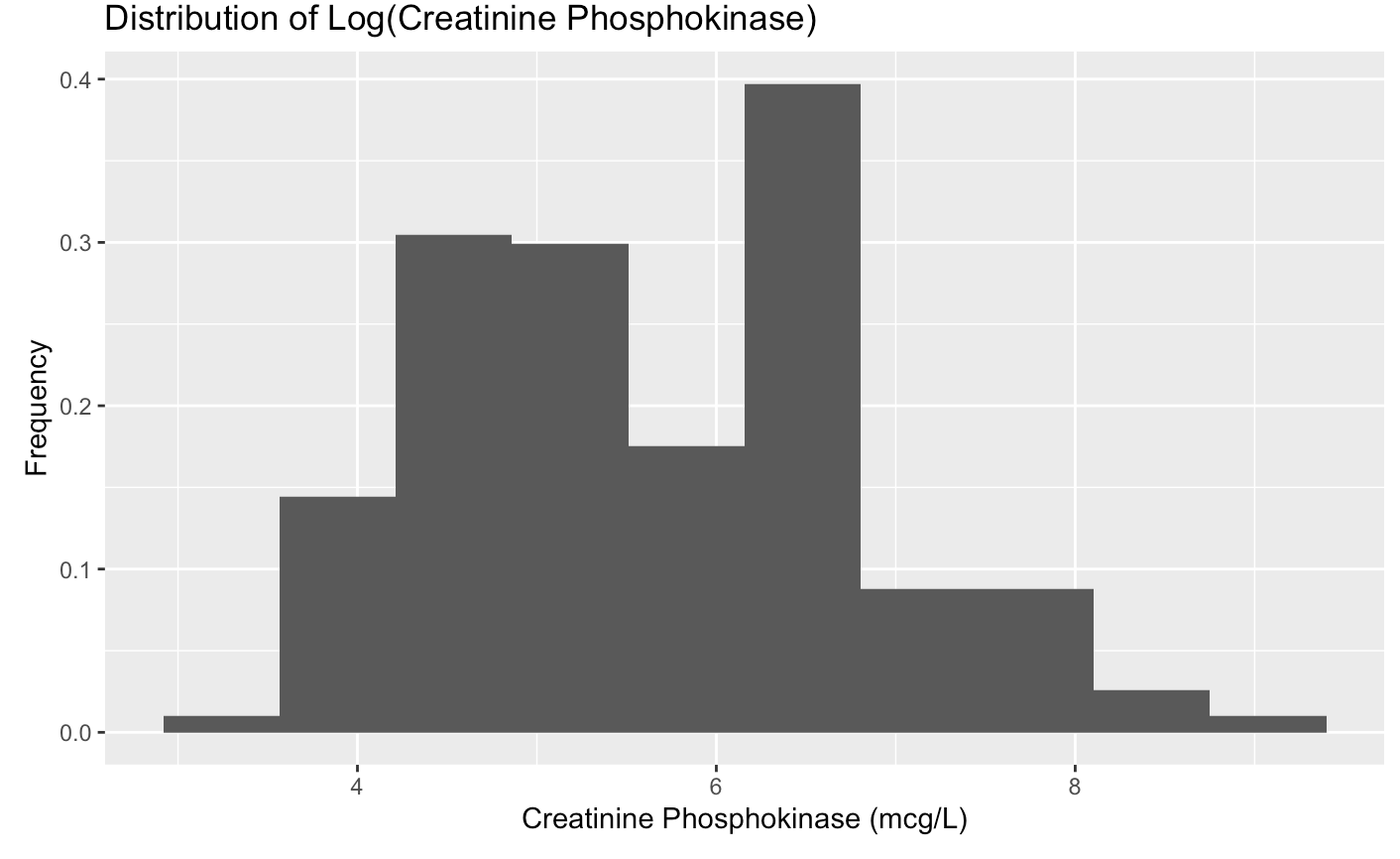
**2. EXPLORATORY DATA ANALYSIS**

1. Data Cleaning

In order to produce the graphical summaries below, we first needed to perform some data cleaning and processing. Firstly, we decided to remove one of the thirteen variables from the original dataset, “time.” This variable describes the follow-up period for the patient, in which they either died or were seen again by their doctor. This variable is not useful for our model, as there is no way to predict the time of death of a patient until after the fact. Next, we found that R read in the binary, categorical variables as numeric variables, so we converted each of these to factor. We also decided to add more descriptive names for each level of the categorical variables. For example, the variable “diabetes” had levels 0 and 1 which we changed to “Diabetic” and “Non-Diabetic” for the sake of clarity. Lastly, we found that the response variable, “creatinine\_phosphokinase”, was heavily skewed to the right (as shown in Figure 1a), so we decided to use a log transform on this variable to even out the distribution. Finally, as with any statistical learning methodology, we needed to split our data into a training data set (in order to train our model) and a test data set (in order to test the trained model). We set a seed for replicability, and then split the data randomly into two equal groups. At this point, we were ready to proceed with the exploratory data analysis, which we performed only on the training data.

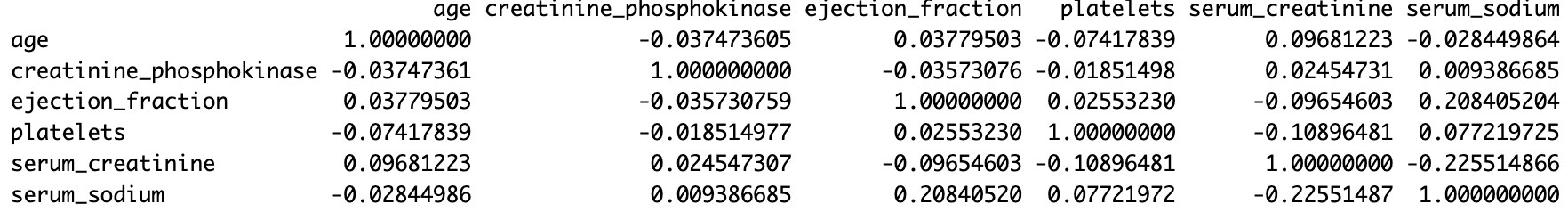
1. Graphical Summaries

***Figure 1A: Figure 1B:***

*** ***

From the 1A, the original response variable, ‘creatinine\_phosphokinase,’ is heavily skewed to the right. In order to correct for this, we transformed the variable using the log function, producing a much more symmetric, albeit not perfectly even distribution (Figure 1B).

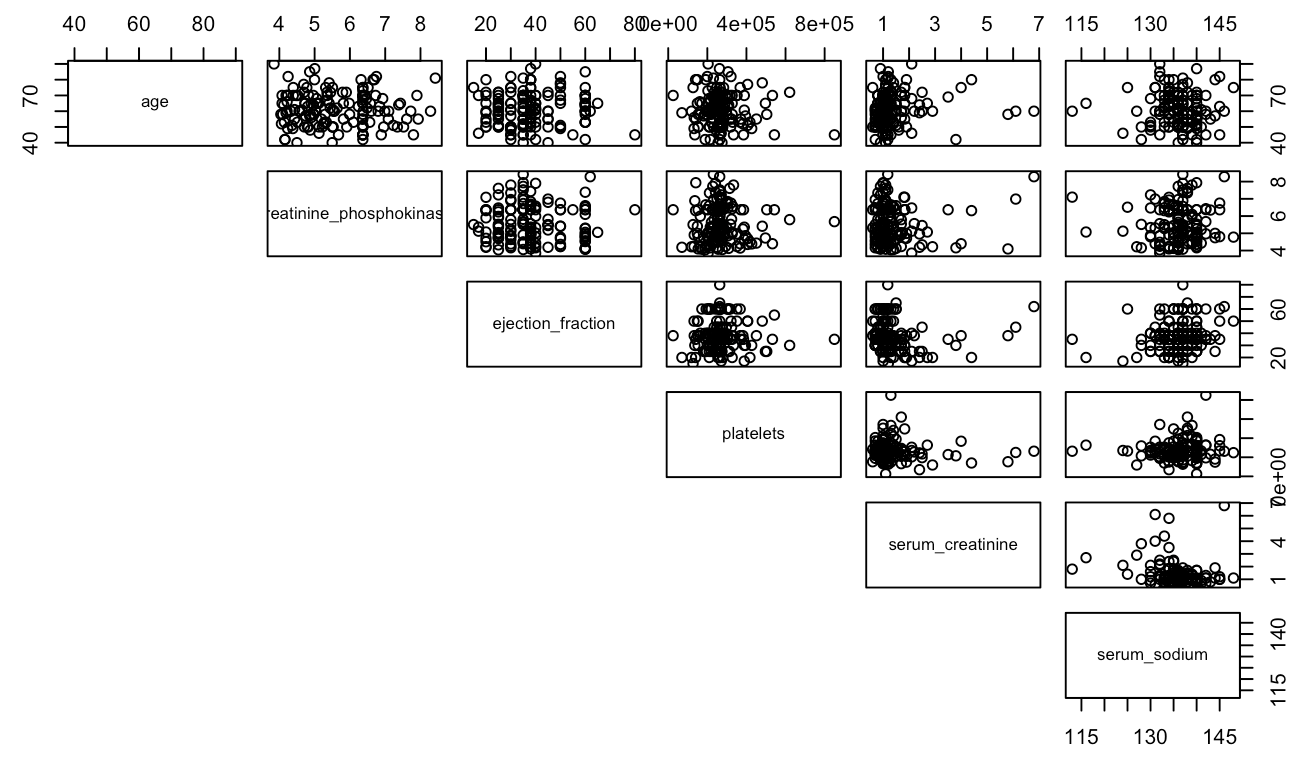
***Figure 2***

******

In the correlation matrix presented above, there are several noteworthy linear relationships between the variables. To begin, age exhibits a weak negative correlation with creatinine phosphokinase (-0.0375) and platelets (-0.0742), indicating almost no correlation with creatinine phosphokinase and a slight tendency for platelet levels to decrease with age. Moreover, in addition to these two negative associations, age is positively correlated with ejection fraction and serum creatinine, suggesting that these variables tend to increase with age.

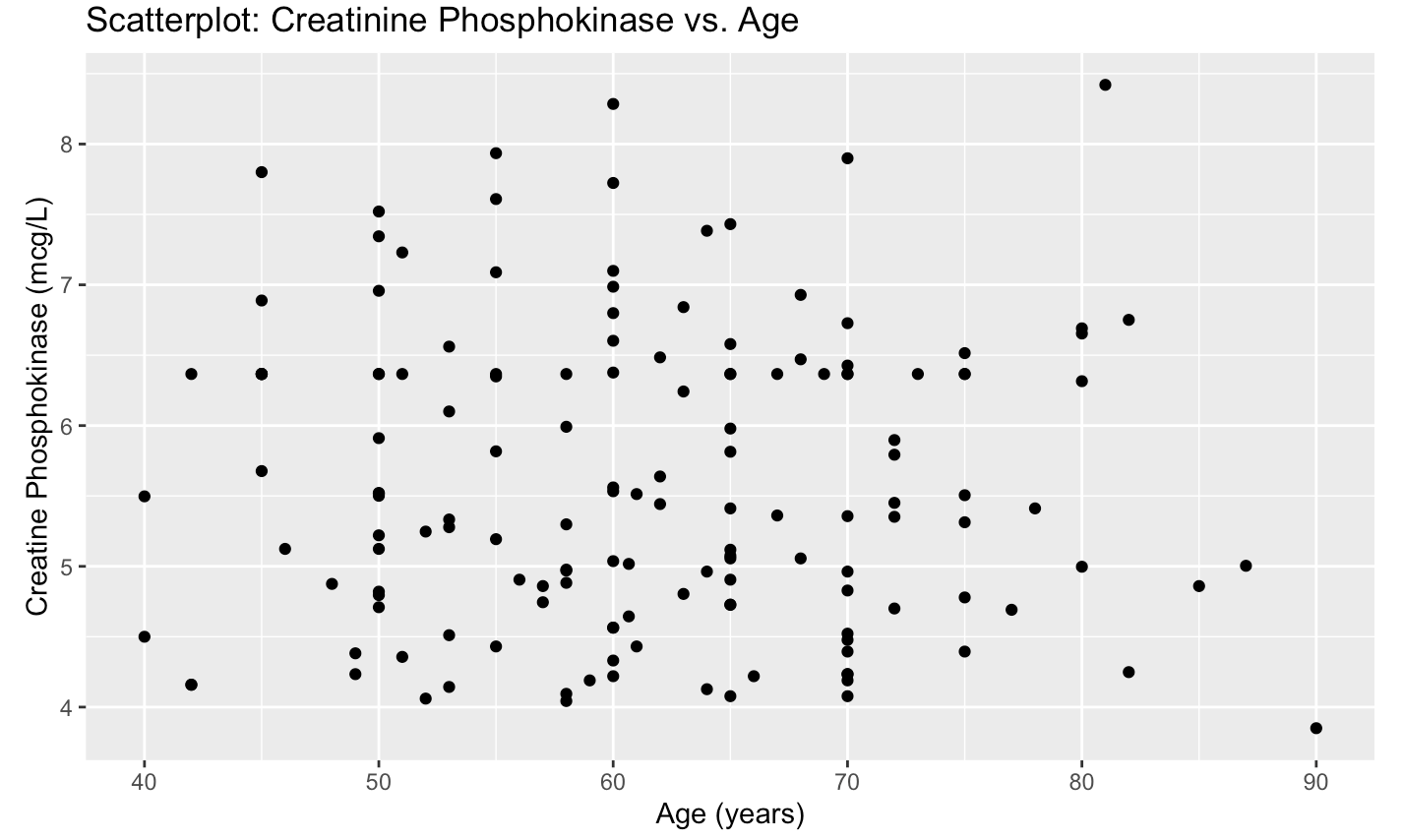
Creatinine phosphokinase, on the other hand, has a weak negative correlation (-0.0185) with platelets, implying a slight decrease in platelet levels as creatinine phosphokinase levels increase. Lastly, there is a weak negative correlation (-0.2255) between serum creatinine and serum sodium, signifying a tendency for serum sodium levels to decrease as serum creatinine levels increase. Furthermore, serum sodium and ejection fraction exhibit a weak positive correlation (0.2084), suggesting that serum sodium levels tend to rise as ejection fraction increases. As none of the variables are strongly correlated with one another, we can conclude that multicollinearity is likely not a significant issue for this dataset.

***Figure 3***



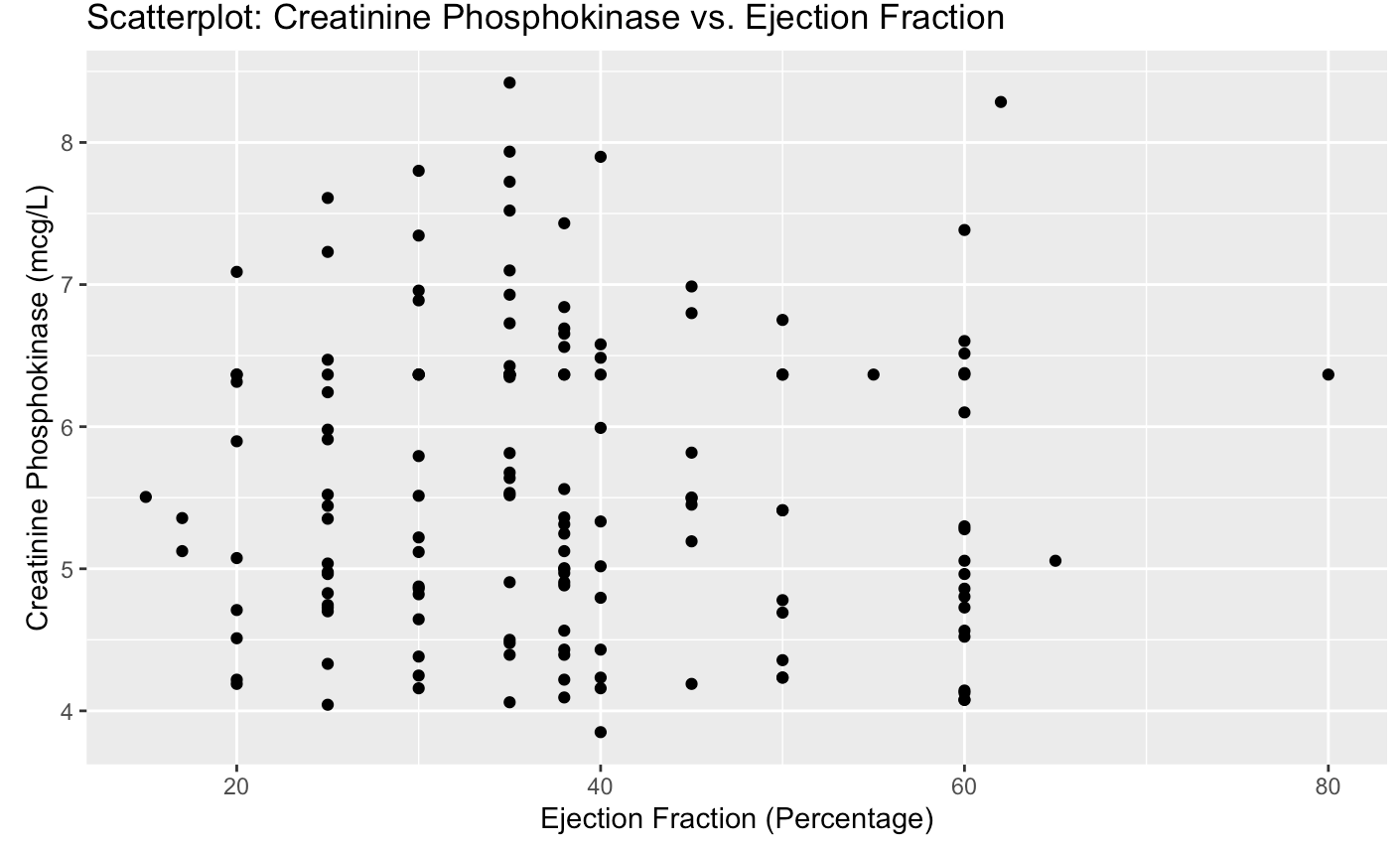
Examining the provided scatterplot matrix, we can see that the majority of relationships between the quantitative variables are very weak. Serum\_sodium exhibits two positive associations, one with platelets and the other with ejection\_fraction. This implies that as serum\_sodium increases, both platelets and ejection\_fraction tend to increase. Additionally, one of the most prominent negative relationships in the matrix is observed between serum\_creatinine and platelets, signifying that as serum creatinine levels rise, platelet counts tend to decrease. In a broader context, the scatterplot matrix reveals that the majority of these predictors appear to have no significant relationships, and are largely independent of each other.

***Figure 4***

******

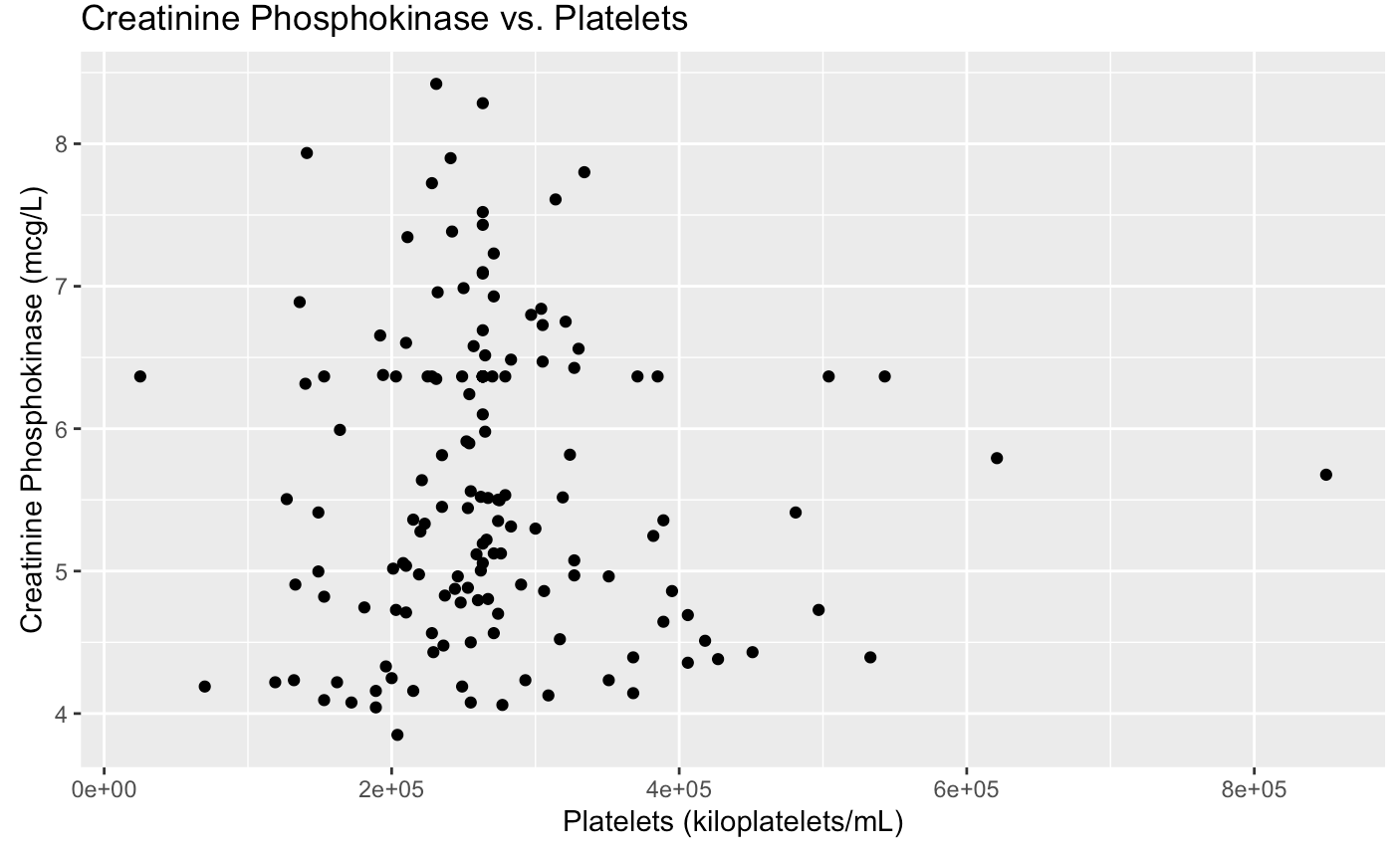
Looking at the scatterplot above, we can see that the relationship between Age and Creatinine Phosphokinase is very weak. With a correlation of -0.0375 (Figure 2), we know that a very weak negative correlation exists, however the points are extremely scattered and do not display a clear trend.

***Figure 5***

******

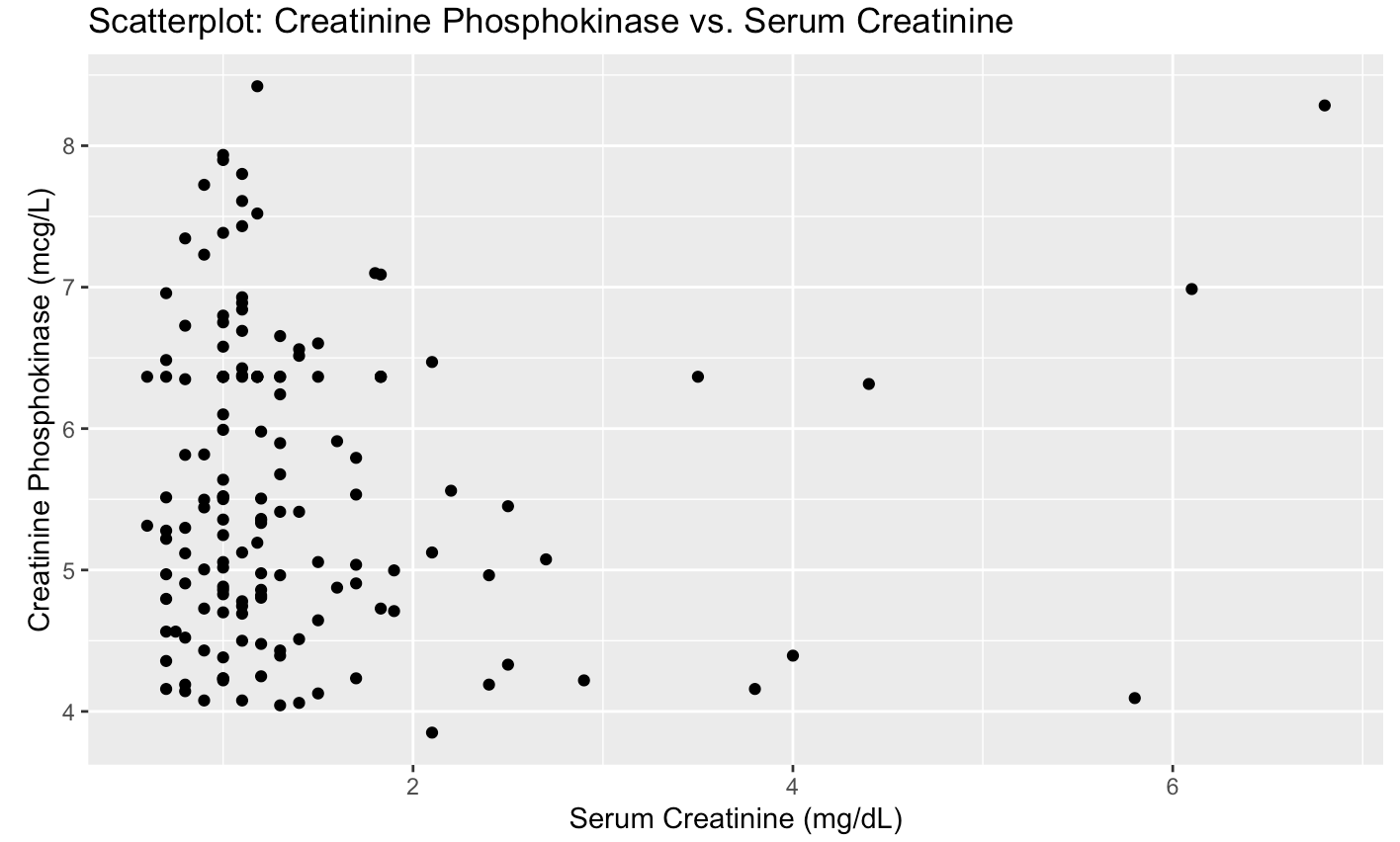
Examining the scatterplot above, we can see that there is not a clear, identifiable relationship present between ejection\_fraction and creatinine\_phosphokinase. The scatterplot does illustrate a handful of outliers that lie at the intersection of low ejection fraction and high CPK levels. These outliers may be a result of the low ejection fraction combining with other personal demographics and health characteristics to produce high CPK enzyme levels, but they could also be a result of natural variation within the population of individuals sampled. Given the weak correlation of -0.0357 (Figure 2), we can conclude that the relationship between these variables is likely negligible.

***Figure 6***



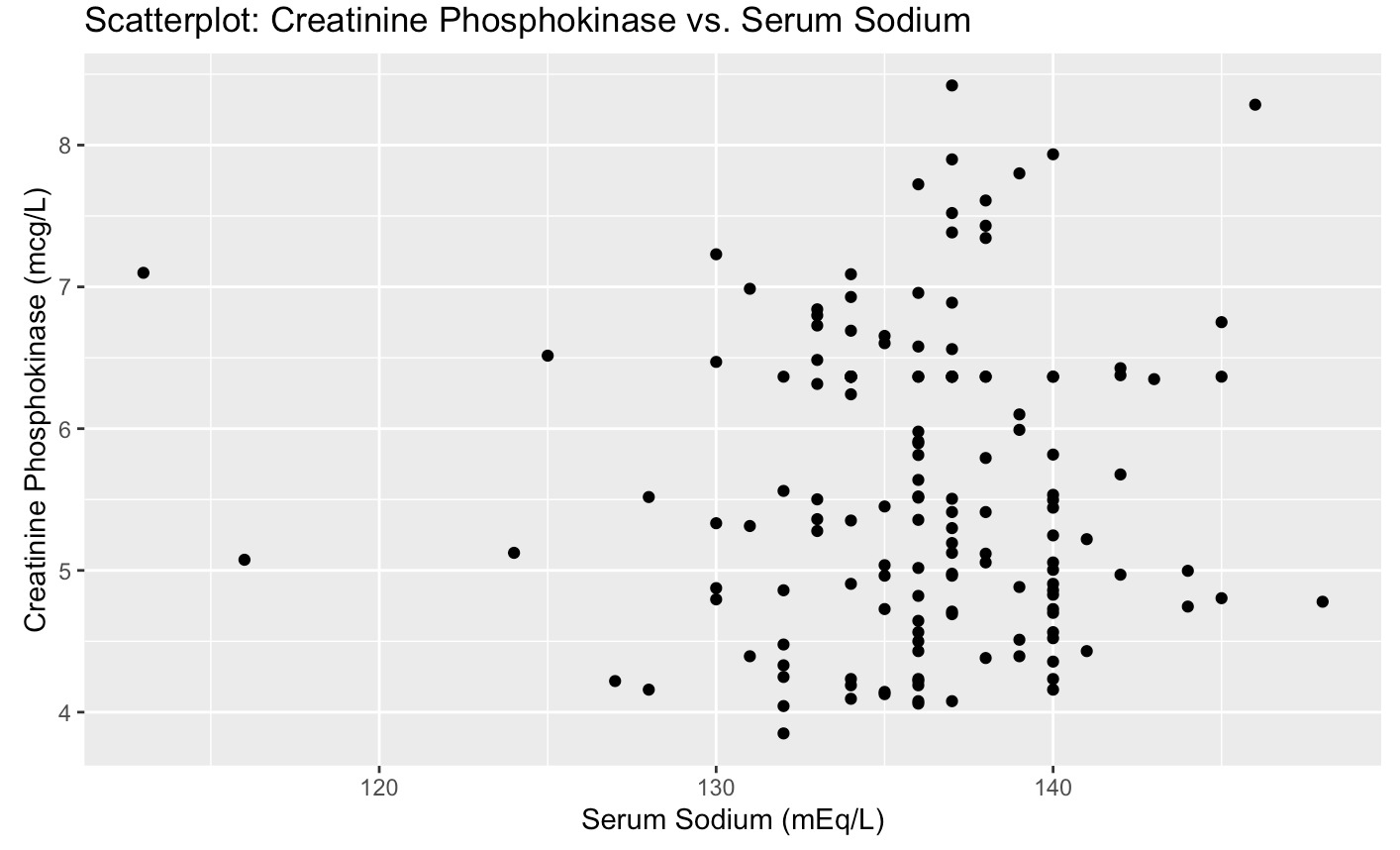
In the above scatterplot, we analyze the relationship between Platelets and Creatinine Phosphokinase. A slightly negative correlation is evident, as a majority of the points display a downward trend. This trend indicates that as platelet levels increase, there is a tendency for creatinine phosphokinase levels to decrease. This association suggests that low platelet levels may pose an increased risk for higher creatinine phosphokinase levels, potentially linked to eventual heart failure.

***Figure 7***

******

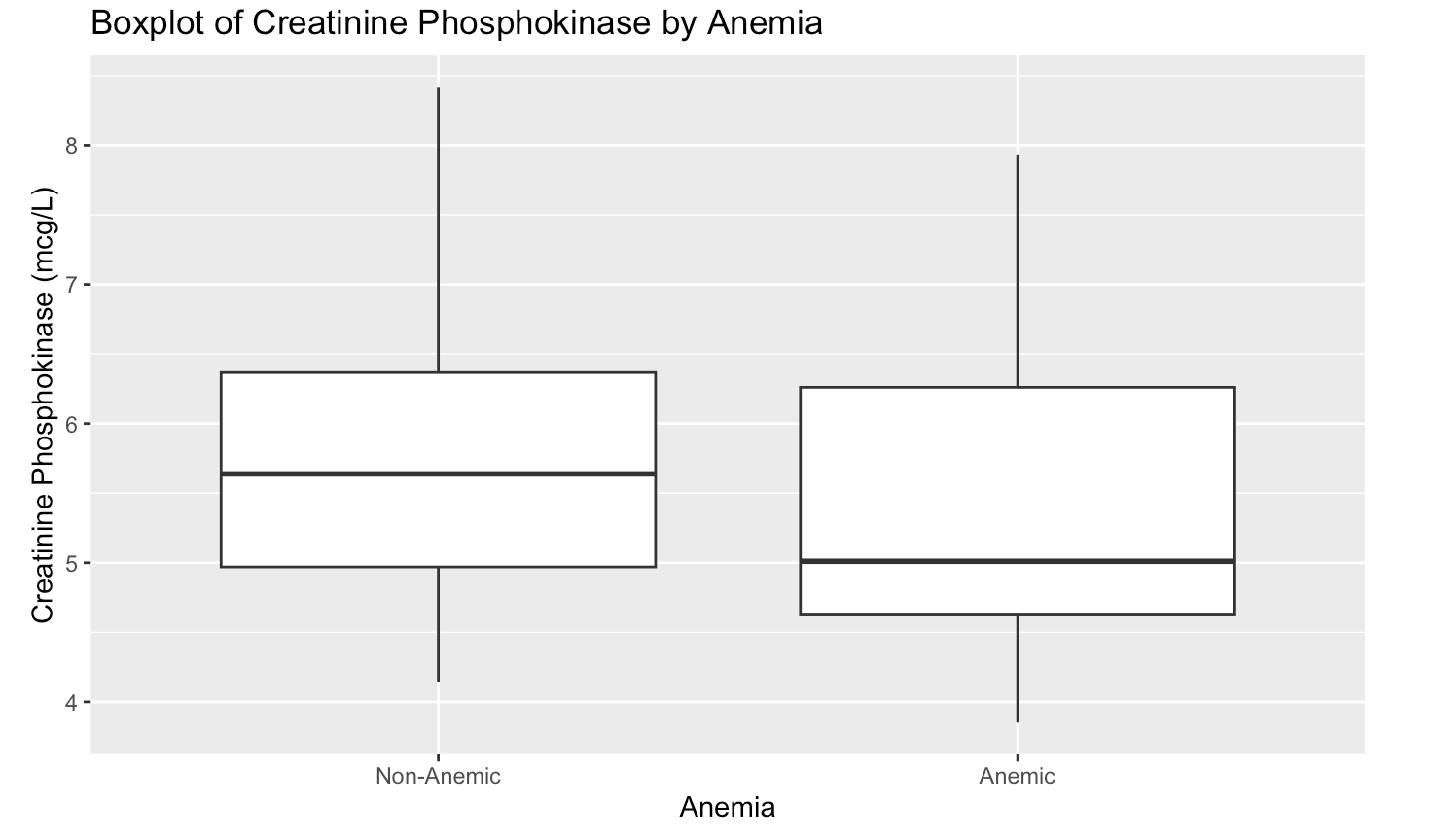
In this scatterplot, it is apparent that there is an absence of a significant relationship between creatinine phosphokinase and serum creatinine levels. The majority of points are densely clustered in the lower left corner, suggesting that creatinine levels tend to remain relatively constant across varying serum creatinine levels. Notably, the plot showcases outliers where creatinine levels remain steady in tandem with serum creatinine, while in other cases, creatinine phosphokinase levels increase while certain serum creatinine levels remain unchanged. This disparity confirms the lack of a strong correlation between the two variables.

***Figure 8***



Observing the presented scatterplot, a slight positive relationship between serum sodium levels and creatinine phosphokinase becomes evident. This relationship suggests that elevations in serum sodium levels correspond, to a limited extent, with increases in creatinine phosphokinase. Such a correlation may signify potential heart damage, raising concerns about the likelihood of heart failure, as heightened sodium levels could contribute to increased cardiac damage.

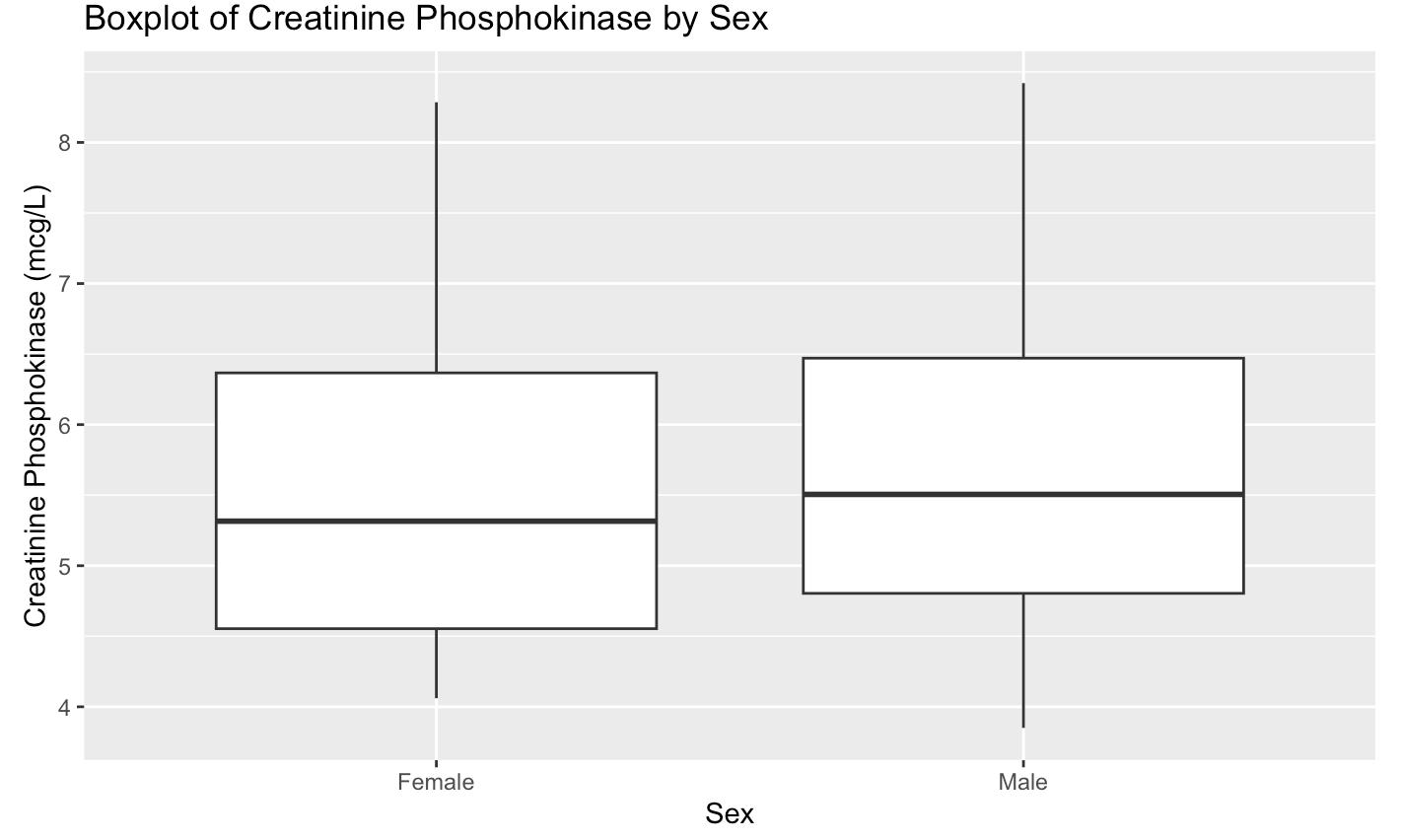
***Figure 9***

******

This boxplot displays the difference in CPK levels between anemic and non-anemic patients. The median value is higher for the first group than the second. This indicates that non-anemic people have higher median CPK levels than anemic people.

The non-anemic group also has more outliers, showing more extreme cases of high CPK concentration. This is surprising because it is typically understood that those that are anemic are likely to have higher levels of CPK, as high creatinine levels indicate issues with kidney function, potentially leading to anemia.

***Figure 10***

******

This graph indicates that men and women are very similar regarding median CPK levels, with men having slightly higher levels on average. This was not surprising, as men on average have a higher percentage of muscle mass than women. The box plot also shows that men have a slightly higher interquartile range (IQR) than women. This indicates that there is more variability in CPK levels among men.

**3. SHRINKAGE METHODS**

1. Data Cleaning

The data cleaning process for the shrinkage methods involved handling the categorical variables that are viewed as factors in R. Unfortunately, the *glmnet* package cannot work with factors, so we converted them to dummy codes. Additionally, as we already split our data into training and test sets, we added vectors that store the predictor and response variables for the test and training data sets. All other relevant data cleaning had already been performed prior to this point in the analysis.

1. Included Predictors

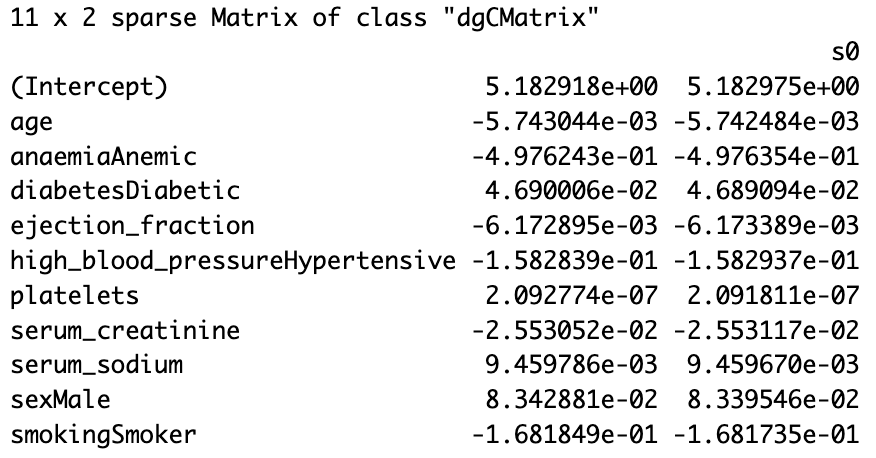
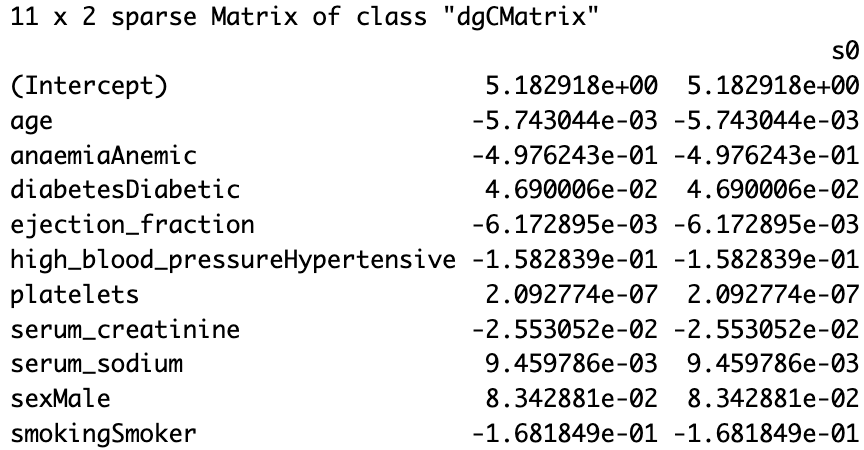
As highlighted in the exploratory data analysis, none of the variables are strongly or directly related to one another, so there was no need to exclude variables as a result of multicollinearity. That being said, one of the variables, DEATH\_EVENT, was excluded, as it does not make sense in the context of the question. We do not know whether someone will die until after the fact, making it a useless predictor in the context of this regression analysis.

|  |  |
| --- | --- |
| Age: | The risk of heart disease increases with age, as the heart muscle can weaken and arteries can stiffen. (U.S. Department) |
| Anaemia | Affects blood volume which disrupts the proper heart function. (Anemia) |
| Diabetes | Diabetes, and the high blood sugar that results, can damage the heart and blood vessels. (Centers) |
| Ejection\_fraction | A lower ejection fraction can indicate that the heart is not effectively pumping blood and is potentially damaged. (Mayo) |
| High\_blood\_pressure | Can add stress to the heart and potentially lead to damage. (Mayo) |
| Platelets | Abnormal levels of platelets can indicate heart disease (2022 AHA). |
| Serum\_creatinine | Can indicate impaired kidney function and lead to heart disease. (Chen) |
| Serum\_sodium | High serum sodium levels can cause an excess retention of water and put more pressure on the heart as it pumps blood. (Mayo) |
| Sex | Men tend to have a higher risk of heart disease, and thus muscle damage than women. (Bots) |
| Smoking | Smoking can increase damage to the heart’s walls. (US Department) |

1. Threshold Value

When choosing the appropriate threshold value for the minimization procedure, it is important that the estimated coefficients for ridge regression are very close to, or ideally equal to, the estimated coefficients from ordinary least squares. This is because when the tuning parameter (λ) is equal to 0, the coefficients should be the same. In order to ensure equality in these coefficients, we first used the default threshold, finding that the coefficients were not equal. Thus, we decided to reduce the threshold to 10-23, which in turn gave us estimated ridge regression coefficients that are equal to the estimated ordinary least squares coefficients.

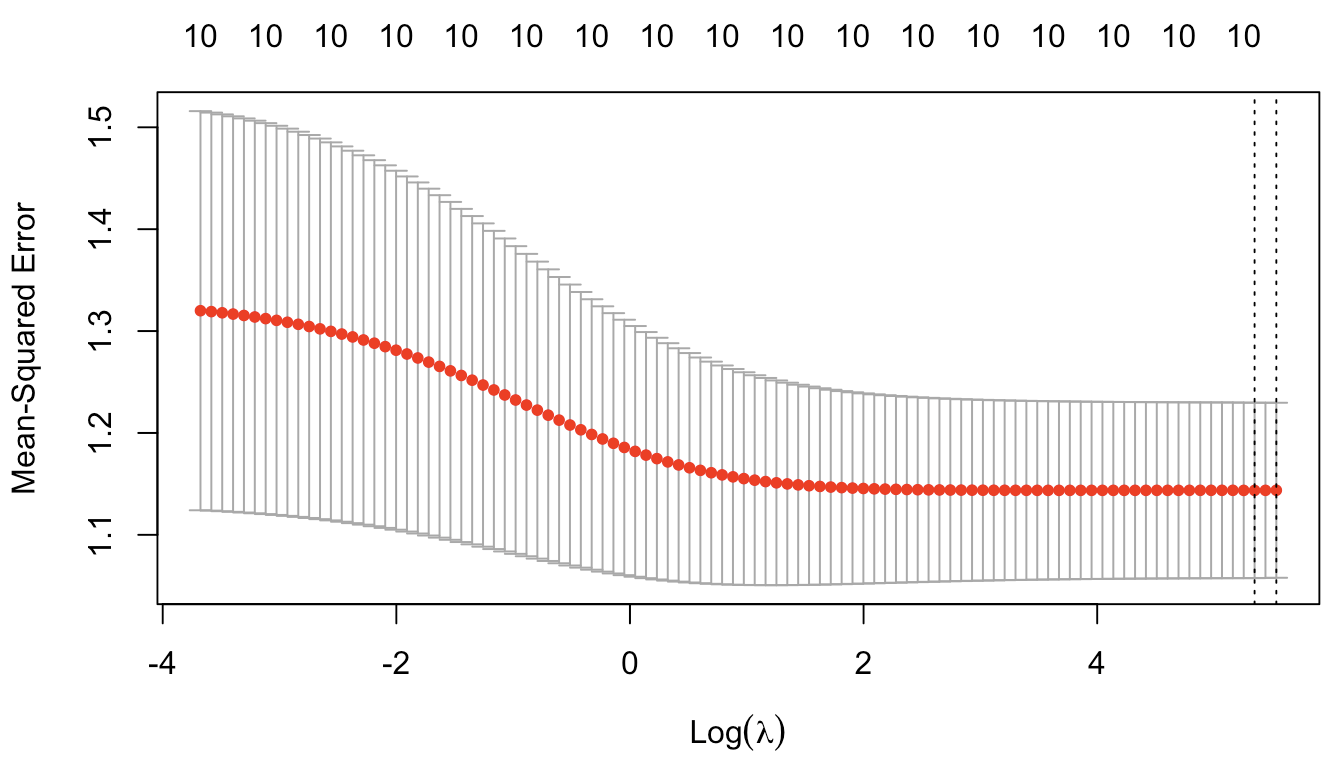
**Threshold: 10-7 Threshold: 10-23**

1. Ridge Regression
   1. **Tuning Parameter**

The value of the tuning parameter, λ, based on 10-fold cross-validation on the training data is **210.0792**

* 1. **Test MSE vs. Log(λ)**



* 1. **Number of Predictors**

The number of predictors that are left in the model based on the value of λ chosen by cross validation is **10**. This is not surprising, as ridge regression keeps all of the predictors.

* 1. **List of Predictors**

age, anaemia, diabetes, ejection\_fraction, high\_blood\_pressure, platelets, serum\_creatinine, serum\_sodium, sex, smoking

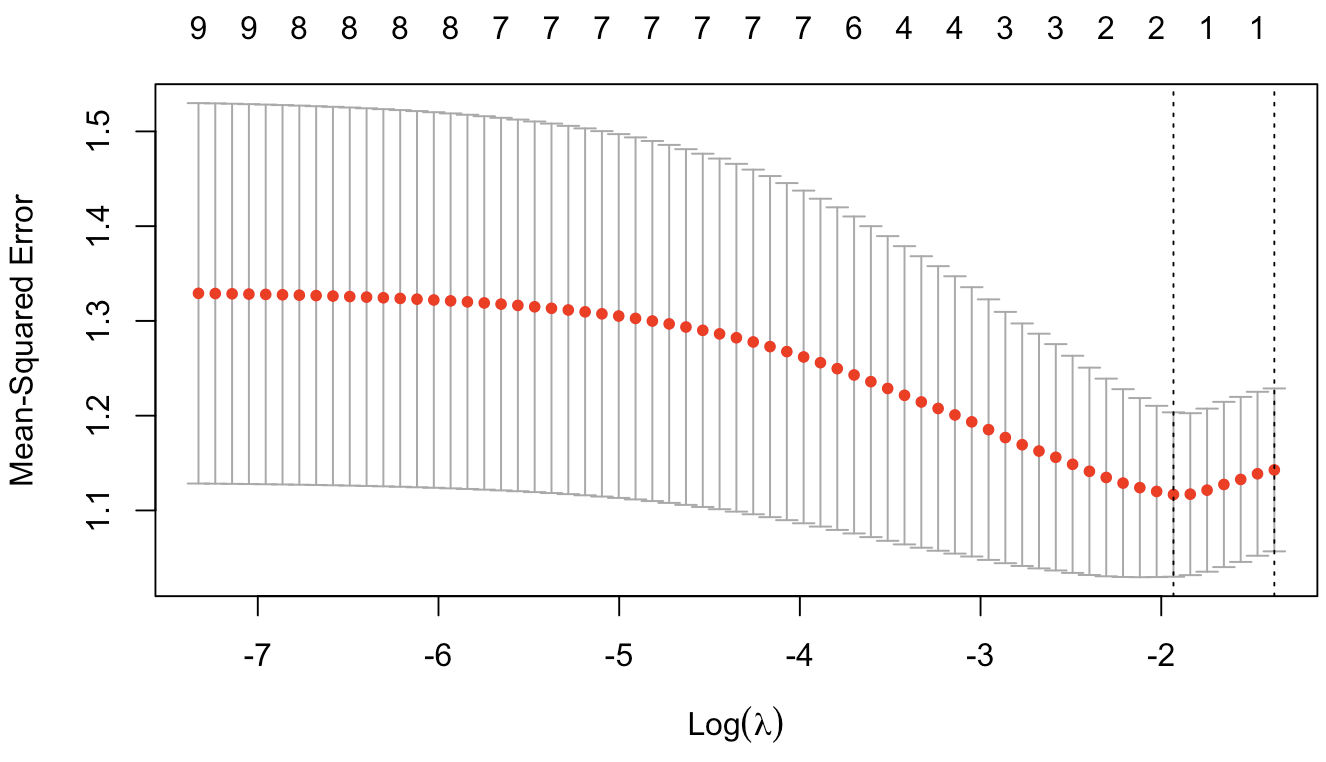
* 1. **Actual Test MSE**

The actual test MSE based on the model using the value of λ chosen by cross validation is: **1.444791**

1. Lasso Regression
   1. **Tuning Parameter**

The value of the tuning parameter, λ, based on 10-fold cross-validation on the training data is **0.1447994.**

* 1. **Test MSE vs. Log(λ)**



* 1. **Number of Predictors**

The number of predictors that are left in the model based on the value of λ chosen by cross validation is **2**. This is not surprising, as Lasso performs variable selection and removes predictors from the model.

* 1. **List of Predictors**

Anemia, sex

* 1. **Actual Test MSE**

The actual test MSE based on the model using the value of λ chosen by cross validation is: **1.402563**

1. Ordinary Least Squares Regression
2. **Actual Test MSE**

The actual test MSE based on the model using the value of λ chosen by cross validation is: **1.438754**

1. Conclusion
   1. **Table: MSEs**

|  |  |  |
| --- | --- | --- |
| **Ridge Regression** | **Lasso Regression** | **OLS Regression** |
| 1.444791 | 1.402563 | 1.438754 |

* 1. **Question of Interest:**

Each of the three methods are useful in analyzing our question of interest which asks about the relationship between a patient’s level of the CPK enzyme (creatinine phosphokinase) and their personal demographics and health history. First, ridge regression can be used to identify which patient attributes are most significant in predicting CPK enzyme level. The regression will shrink the less important predictors but still retain them with reduced impact through regularization. Secondly, the lasso regression can be used for variable selection, helping to identify which variables are most related to CPK enzyme levels. It differs from ridge regression, as all predictors are retained, but their impact is proportionally reduced. Finally, OLS regression can be used to determine the relationship between CPK enzyme levels and the variables without any form of regularization by offering direct estimates of each coefficient.

* 1. **Discussion**

Looking at our three regression models, we can conclude that the ridge regression model is the best-performing model in predicting CPK enzyme level. The Lasso regression has the lowest test MSE (1.402563) compared to Ridge (1.444791) and OLS (1.438754). This low test MSE implies that the lasso regression could provide the closest predictions to the actual CPK enzyme levels. One of the possible reasons for this model being the best predicting CPK levels is the fact that lasso regression can handle multicollinearity by driving some of the coefficients to zero. Another possibility is that the Lasso regression performs better when there are only a few predictors that truly impact the response variable, which could be the case in this dataset.

* 1. **Challenges**

One of the main challenges we faced was determining the most relevant predictors that can be used to predict CPK enzyme level. We did not want to use too many predictors because that could lead to overfitting of the model which could increase the variance of the mean prediction.

**4. REGRESSION TREE**

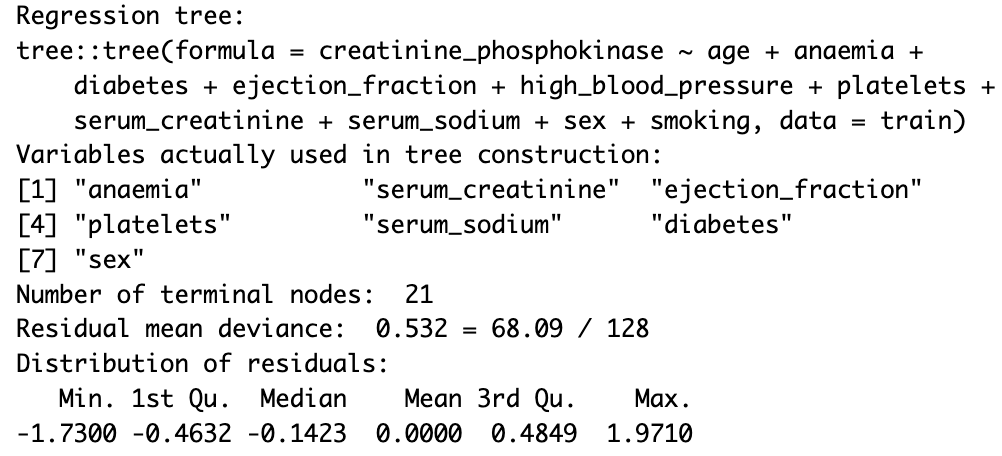
1. Data Cleaning

The first part of the data cleaning process for the regression tree portion of our analysis is to ensure that our categorical variables are stored as factors, as the *tree* package cannot handle dummy codes. Additionally, it is important to have a vector that stores the response variable for the test data. Both of these steps were already taken care of in the beginning portions of our project, and therefore no further cleaning was needed.

1. Included Predictors

Refer to Part 3b (“Included Predictors”).

1. Regression Tree with Recursive Binary Splitting
   1. **Summary Output**



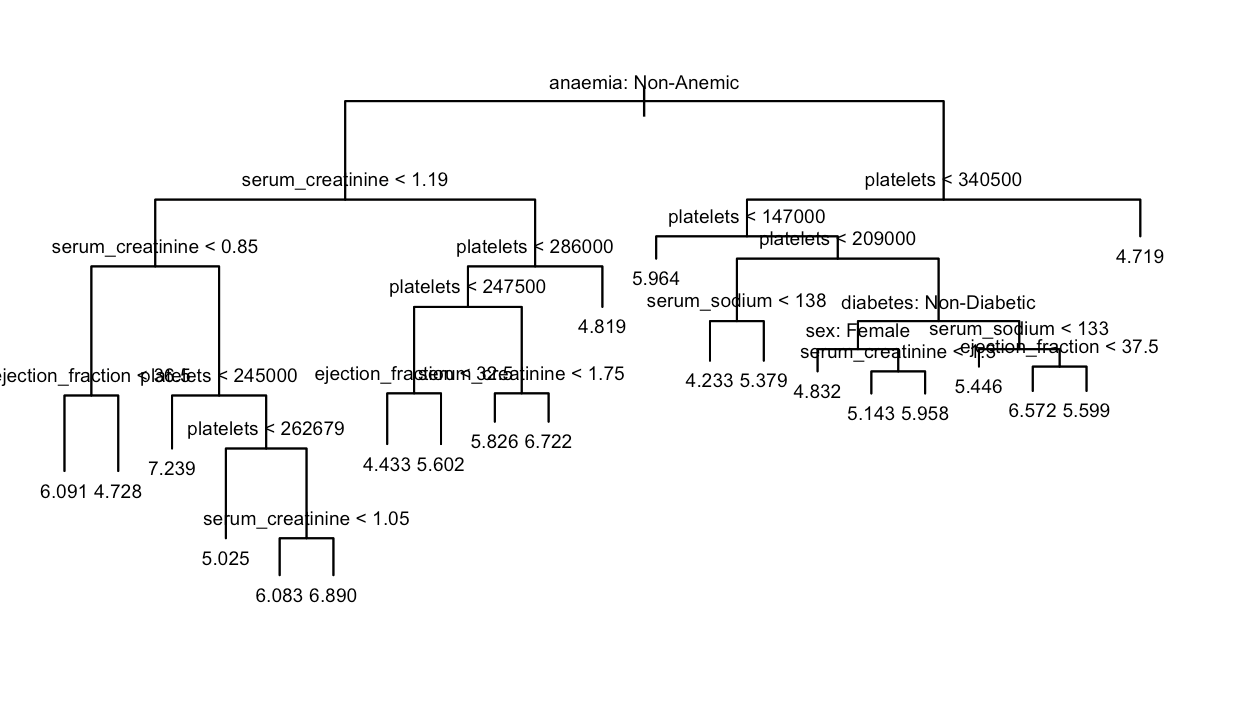
* 1. **Terminal Nodes**

The tree has **21** terminal nodes.

* 1. **Predictors**

anaemia, diabetes, ejection\_fraction, platelets, serum\_creatinine, serum\_sodium, serum\_creatinine, sex

* 1. **Graphical Output**



* 1. **Question of Interest**

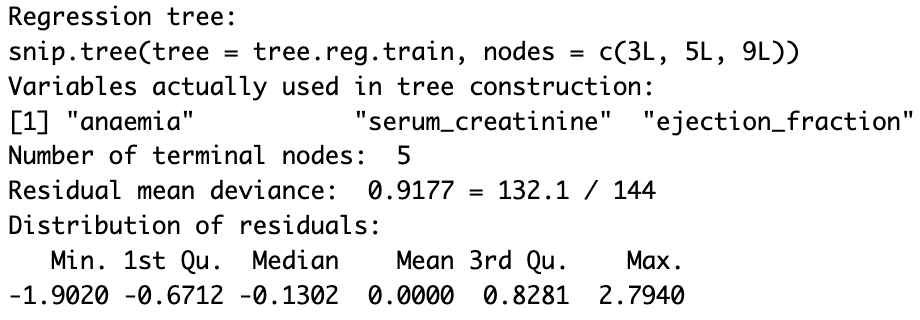
The regression tree effectively addresses our question regarding CPK enzyme levels. It begins with an initial node based on the presence of anemia in the patient, leading to further divisions into decision nodes. These nodes, which branch further into additional decision nodes or terminal nodes, offer pivotal insights into our area of interest. Through the values of variables within each node, we determine if an individual conforms to certain conditions, thereby determining the anticipated CPK enzyme level. For instance, a patient without anemia, exhibiting serum creatinine levels below 1.19 and 0.85, alongside platelet counts under 245,000, is predicted to have a log CPK enzyme level of 7.239, which when exponentiated, is 1,392.7 mcg/L. This predictive value is based on the average observation within the training set, characterizing that specific segment of the data.

These predictors were chosen for this decision tree because they all could have a strong impact on the response variable. The presence of anemia could have a substantial impact on physiological functions and is associated with changes in CPK enzymes. Deviations from the normal serum sodium levels cause the CPK enzyme levels to increase. Lastly, having a higher platelet level is associated with thrombocytosis and thrombocythemia, which can cause higher CPK levels. (National Heart, Lung, and Blood Institute)

* 1. **Test MSE**

The test MSE from this regression tree is 1.860219

1. Pruned Regression Tree
   1. **Summary Output**



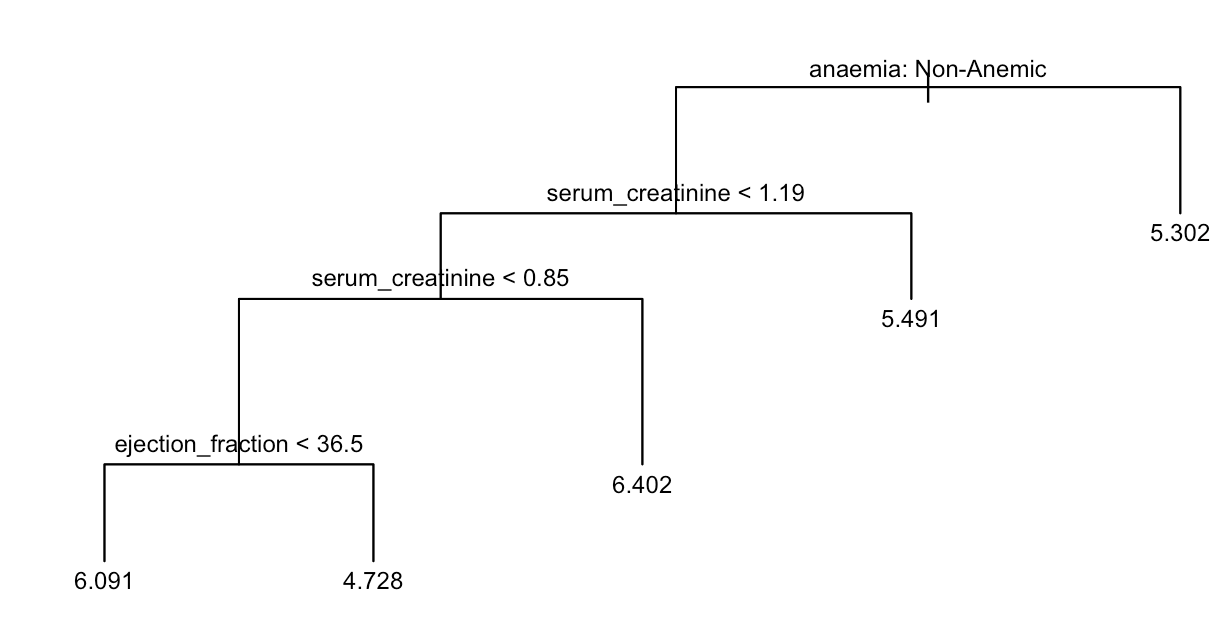
* 1. **Terminal Nodes**

The tree has 5 terminal nodes.

* 1. **Predictors**

The predictors used in this regression tree were anemia, serum\_sodium, DEATH\_EVENT, platelets

* 1. **Graphical output**



* 1. **Question of Interest**

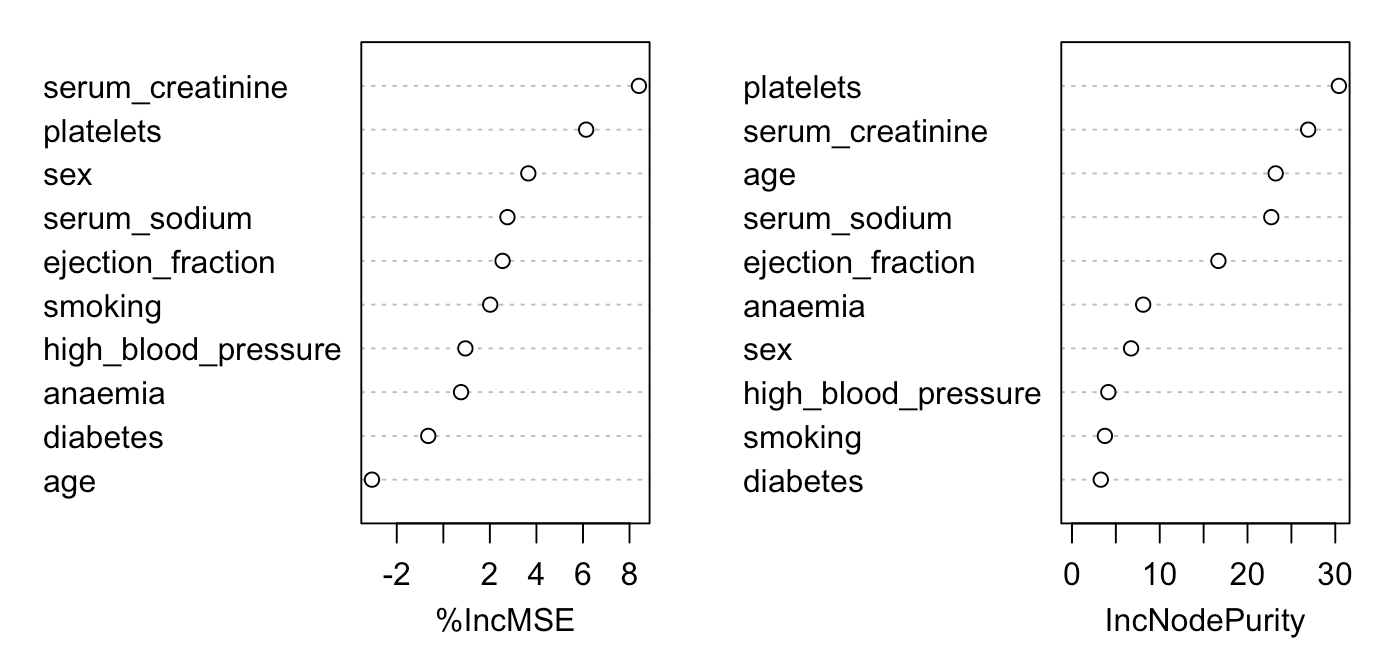
The regression tree presented serves as a beneficial tool for exploring the relationship with CPK enzyme levels. It originates from the root node based on the presence of anemia, branching into decision nodes delineated by serum creatinine and ejection fraction variables. These divisions identify specific conditions that influence the expected CPK enzyme level for individuals. For instance, a non-anemic patient with a serum creatinine level between 0.85 and 1.19 is predicted to have a log CPK enzyme level of 6.402, which when exponentiated is 603.05 mcg/L.

The selection of these predictors within the decision tree is deliberate, chosen for their potential significant impact on the response variable. Anemia, which is known to affect physiological functions, demonstrates an association with changes in CPK enzymes. Deviations from normal serum sodium levels correlate with shifts in CPK enzyme levels Lastly, higher platelet counts, associated with thrombocytosis and thrombocythemia, contribute to elevated CPK levels (National Heart, Lung, and Blood Institute). These variables play important roles in understanding CPK enzyme levels, and the regression tree effectively segments the data based on their values within each node to predict and identify conditions likely to impact CPK enzyme levels.

* 1. **Test MSE**

The test MSE for this pruned tree is 1.487567

1. Tree Improvements using Random Forests
   1. **Important predictors**

****

Based on the improvements to the tree made using random forests we consider the predictors serum\_creatinine, platelets, sex, serum\_sodium, and age to be the most important.

* 1. **Test MSE**

The test MSE for the random forest tree is **1.398362**.

1. Conclusion
   1. **Test MSEs:**

|  |  |  |
| --- | --- | --- |
| **Recursive Binary Splitting** | **Pruned Regression Tree** | **Random Forest** |
| 1.860219 | 1.487567 | 1.398362 |

* 1. **Question of Interest**

These models contribute to understanding how various predictors impact CPK enzyme levels and enhance predictive accuracy. The trees generated using the recursive binary splitting and the pruned method create conditions based on predictors to predict different CPK levels. Analyzing these branching conditions helps us understand how changes in predictors correspond to CPK enzyme level variations. The random forest model takes insights from both decision trees and computes the importance of predictors in influencing CPK enzyme levels.

* 1. **Discussion**

Upon evaluating our three models, it becomes evident that the random forest model outperforms in accurately predicting CPK enzyme levels based on our specific predictors. Comparing the Test MSE values, the Random Forest exhibited the lowest score (1.398362) compared to the recursive binary splitting (1.860219) and the pruned regression tree model (1.487567). The reason for the random forest model achieving heightened accuracy could potentially be attributed to its methodology of aggregating multiple trees trained on varied data subsets, resulting in superior predictive performance.

* 1. **Challenges**

One of the main challenges we faced was determining the most relevant predictors that can be used to predict CPK enzyme level. We did not want to use too many predictors because that could lead to overfitting of the model when doing recursive binary splitting which could increase the variance of the mean prediction. Additionally, we struggled with interpreting the results, as we did not know whether or not the predictors were accurately capturing the relationship between CPK enzyme levels when using random forest.

**5. CLASSIFICATION TREE**

1. **Data Cleaning**

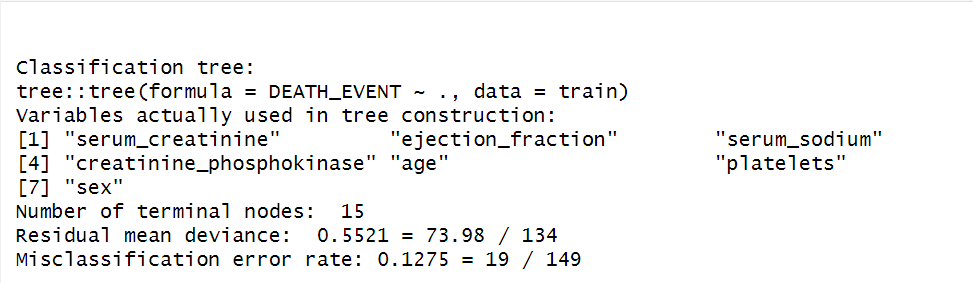
Much of the necessary data cleaning for the classification tree, such as converting categorical variables to factors, was executed in previous steps. We took extra care to ensure that the response variable, DEATH\_EVENT, was considered a factor. Also, now that creatinine\_phosphokinase was no longer the response variable, we reversed the log transformation from a previous step. To do so, we simply exponentiated the variable to bring it back to its original value. We then split the data into training and test groups, and stored the response variable for the test data.

1. **Predictors**

As mentioned previously, none of the variables are strongly or directly related to one another, so there was no need to exclude variables as a result of multicollinearity. We previously dropped the “time” variable in a prior step, as there is no way to predict the time of death of a patient until after the fact.

|  |  |
| --- | --- |
| Age: | The risk of heart disease increases with age, as the heart muscle can weaken and arteries can stiffen (U.S. Department). This may increase the likelihood of a death event. |
| Anaemia | Anaemia ffects blood volume which disrupts the proper heart function and may lead to death. (Anemia) |
| Diabetes | Diabetes, and the high blood sugar that results, can damage the heart and blood vessels, which could lead to death. (Centers) |
| Ejection\_fraction | A lower ejection fraction can indicate that the heart is not effectively pumping blood and is potentially damaged, indicating a health issue ending in death. (Mayo) |
| High\_blood\_pressure | Can add stress to the heart and potentially lead to death or injury. (Mayo) |
| Platelets | Abnormal levels of platelets can indicate heart disease, which many people pass away from. (2022 AHA). |
| Serum\_creatinine | Can indicate impaired kidney function and lead to heart disease, which many people pass away from. (Chen) |
| Serum\_sodium | High serum sodium levels can cause an excess retention of water and put more pressure on the heart as it pumps blood. (Mayo) |
| Sex | Men tend to have a higher risk of heart disease, and thus muscle damage than women, which could lead to health issues. (Bots) |
| Smoking | Smoking can increase damage to the heart’s walls, increasing the chance of death. (US Department) |
| Creatinine\_Phosphokinase | High levels of the CPK enzyme can indicate that there are internal damages to a patient’s brain or muscle tissue, which could cause death from heart failure. (Chen) |

1. **Classification Tree**
   1. **Summary() output**

****

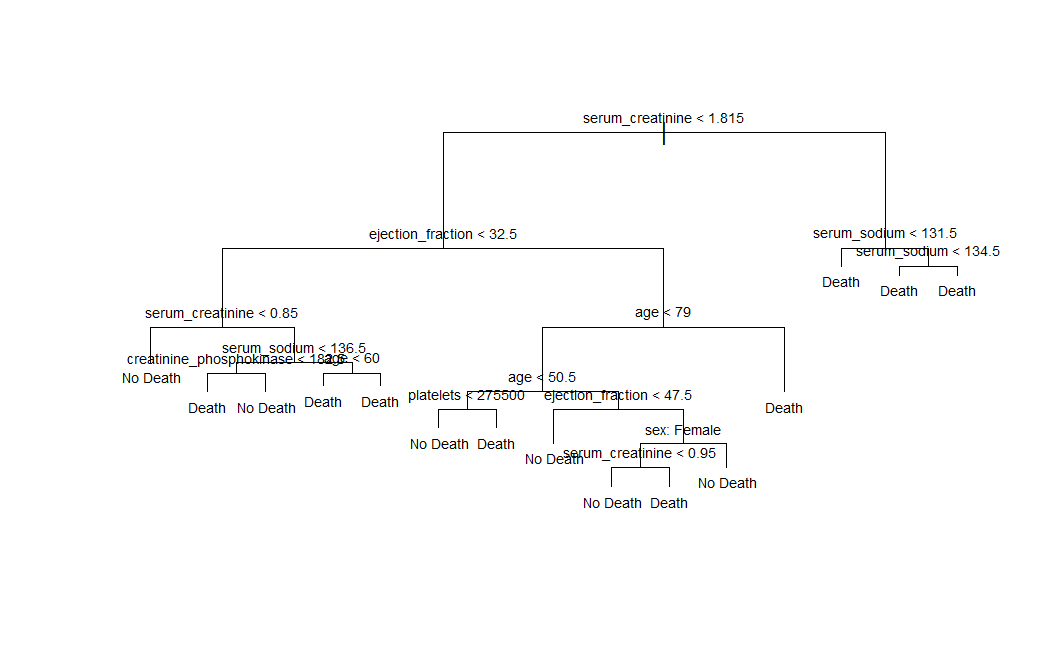
* 1. **Terminal Nodes**

This tree has 15 terminal nodes

* 1. **Predictors**

The predictors used in this tree were serum\_creatinine, ejection\_fraction, serum\_sodium, creatinine\_phosphokinase, age, platelets, and sex.

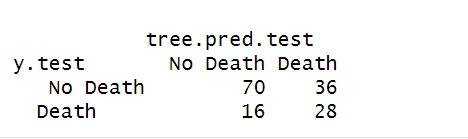
* 1. **Graphical Output**

****

* 1. **Question of Interest**

This classification tree serves to address questions about the predictors associated with individuals dying from heart failure. The root node is centered on serum creatinine levels and separates into multiple decision nodes considering various variables. These distinct conditions on the decision nodes help identify specific variables that are relevant in forecasting an individual's likelihood of experiencing heart failure. For instance, an individual exhibiting a serum creatinine level below 1.815, an ejection fraction surpassing 32.5, and an age exceeding 79 years is predicted to be at risk of heart failure. All predictors play a vital role in predicting an individual's heart failure outcome. Increases in age, serum creatinine, serum sodium, creatinine phosphokinase, and lower ejection fraction levels collectively contribute to heightened probabilities of passing away due to heart failure.

* 1. **Confusion Matrix (Threshold=0.5)**

****

* 1. **Overall Test Error Rate (Threshold=0.5)**

0.3466667

* 1. **False Positive Rate (Threshold=0.5)**

0.3396226

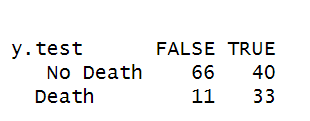
* 1. **False Negative Rate (Threshold=0.5)**

0.3636364

* 1. **Discussion of Threshold**

After reviewing the confusion matrix, we conclude that our false negative rate is far too high. A false negative in a medical context is far more dangerous than a false positive. If a patient is incorrectly predicted to not experience a death event, they will likely miss out on life-saving care. On the other hand, if a patient is incorrectly predicted to die (false positive), there will be monetary loss from wasted medical treatments, but they will not lose their lives. Therefore, to improve our model, it is crucial to reduce the false negative rate (FNR). To do so, we decreased the threshold to 0.30 and produced the following numbers.

* + 1. **Confusion Matrix (Threshold=0.2)**

****

* + 1. **Overall Test Error Rate (Threshold=0.2)**

0.34

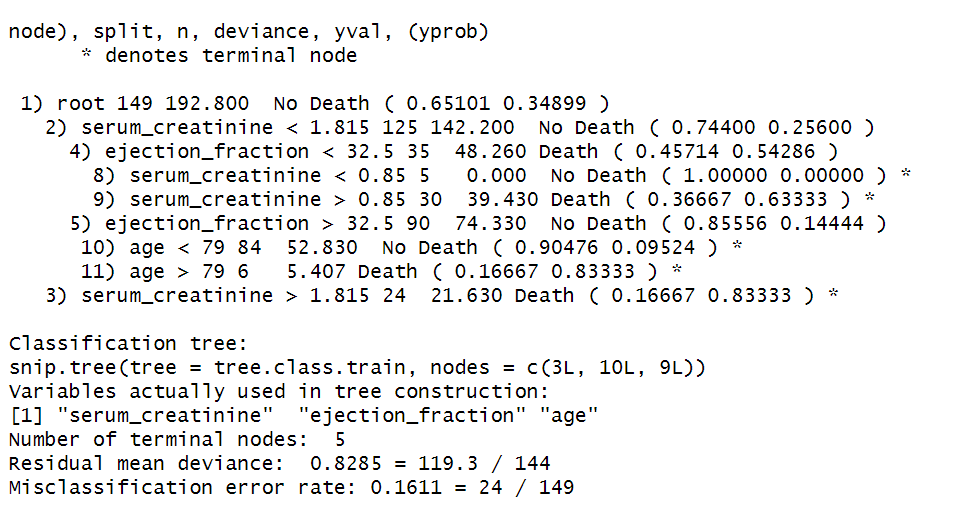
* + 1. **False Positive Rate (Threshold=0.2)**

0.37736

* + 1. **False Negative Rate (Threshold=0.2)**

0.25

1. **Pruned Classification Tree**
   1. **Summary() output**

****

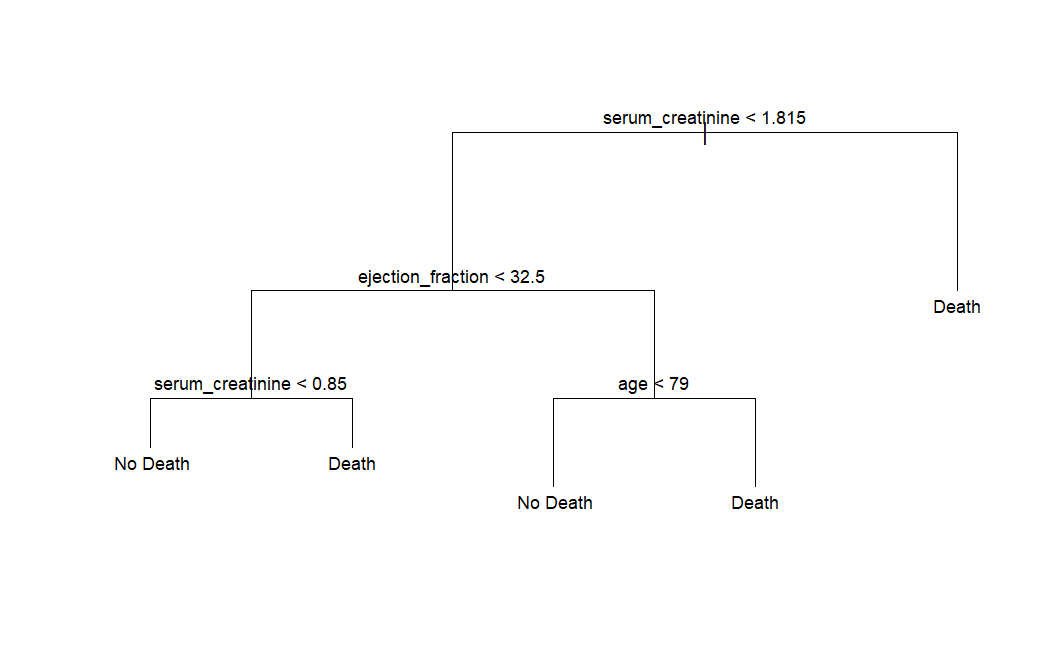
* 1. **Terminal Nodes**

This tree has 5 terminal nodes

* 1. **Predictors:**

The predictors used in this tree were serum\_creatinine, ejection\_fraction, and age.

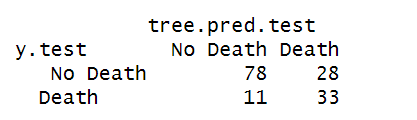
* 1. **Graphical output**



* 1. **Question of Interest**

This classification tree serves as a beneficial tool for analyzing the characteristics that contribute to an individual potentially passing away from heart failure. The root node is based on serum creatinine levels and then divides into a terminal node and a decision node that is based on ejection fraction levels. It is then divided into two further decision nodes based on age and serum creatinine levels again. These conditions help us to determine which values of these predictors indicate whether a patient is likely to die from heart failure. For instance, a patient with serum creatinine levels under 0.85 and ejection fraction levels under 32.5 is not predicted to die from heart failure. All of these predictors are relevant in predicting heart failure, as increases in serum creatinine and age increase the likelihood and decreases in ejection fraction levels increase the likelihood of dying from heart failure.

* 1. **Confusion Matrix**



* 1. **Overall Test Error Rate (Threshold=0.5):**

0.26

* 1. **False Positive Rate (Threshold=0.5):**

0.2641509

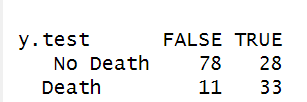
* 1. **False Negative Rate (Threshold=0.5):**

0.25

* 1. **Discussion of Threshold**

Once again, we are concerned with reducing the false negative rate (FNR) in order to avoid loss of life from failure to provide necessary care. This pruned tree has decreased the FNR from 0.3636364 to 0.25 at the threshold of 0.5. However, 0.25 is still higher than necessary, so we planned to decrease the threshold in the pruned example as well. Unfortunately, the confusion matrix numbers did not change with a reduction in threshold up to and including 0.10. Past that point, we risked a large decrease in accuracy so we decided against decreasing the threshold from 0.5.

* + 1. **Confusion Matrix (Threshold = 0.1)**

****

* + 1. **Overall Test Error Rate (Threshold=0.1)**

0.26

* + 1. **False Positive Rate (Threshold=0.1)**

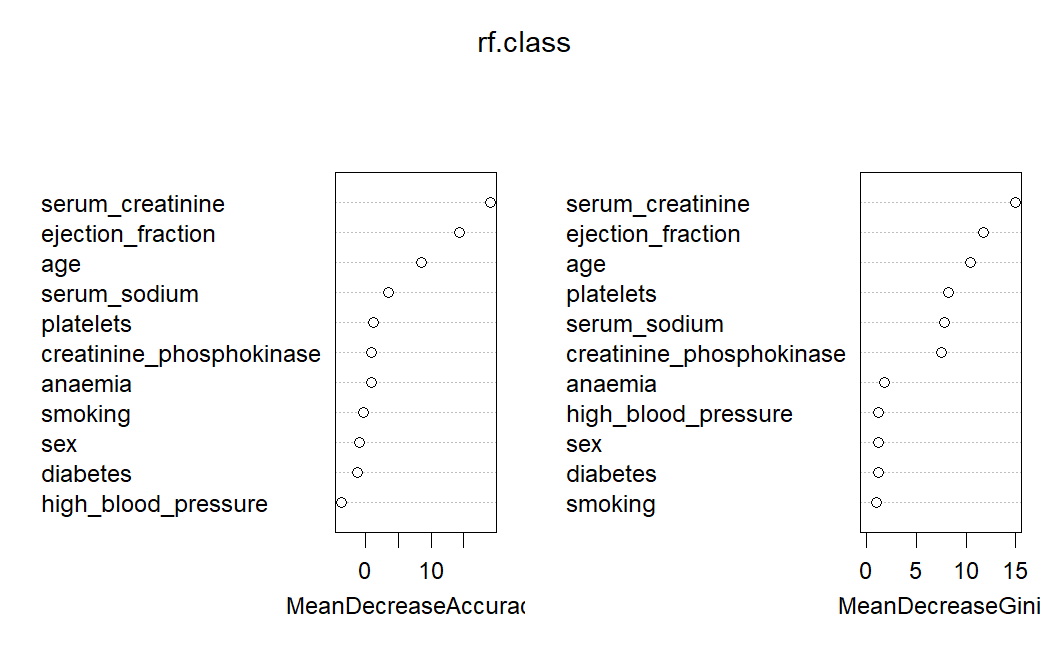
0.2641509

* + 1. **False Negative Rate (Threshold=0.1)**

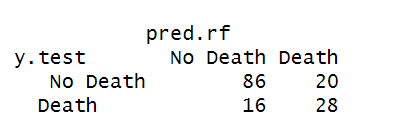
0.25

1. **Random Forests Improvement to Tree**
   1. **Important Predictors:**

The three predictors found most important in random forests are serum\_creatinine, ejection\_fraction, and age.



* 1. **Confusion Matrix (Threshold=0.5)**

****

* 1. **Overall Test Error Rate (Threshold=0.5)**

0.24

* 1. **False Positive Rate (Threshold=0.5)**

0.1886792

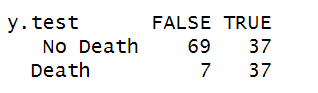
* 1. **False Negative Rate (Threshold=0.5)**

0.3636364

* 1. **Discussion of Threshold:**

Similar to the original recursive binary splitting classification tree, the false negative rate (FNR) is a very high 0.3636364. Once again, we explored the idea of reducing the threshold to prevent deaths from lack of medical attention. This time, however, was much more successful than the attempt with the pruned tree. A reduction in threshold from 0.5 to 0.3 reduced the FNR by more than half, with a new value of 0.15909. Though this increased the overall test error rate slightly and the false positive rate (FPR) greatly, this change is beneficial to our model in that it produced the lowest FNR out of all of our attempts so far.

* + 1. **Confusion Matrix (Threshold = 0.3)**

****

* + 1. **Overall Test Error Rate (Threshold=0.3)**

0.29333

* + 1. **False Positive Rate (Threshold=0.3)**

0.34906

* + 1. **False Negative Rate (Threshold=0.3)**

0.15909

1. **Conclusion:**
2. **Table**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Recursive Binary Splitting** | **Pruned Classification Tree** | **Random Forest** |
| **Test Error Rate** | 0.3466667  *Threshold = 0.5* | 0.26  *Threshold = 0.5* | 0.24  *Threshold = 0.5* |
| **FPR** | 0.3396226  *Threshold = 0.5* | 0.2641509  *Threshold = 0.5* | 0.1886792  *Threshold = 0.5* |
| **FNR** | 0.3636364  *Threshold = 0.5* | 0.25  *Threshold = 0.5* | 0.3636364  *Threshold = 0.5* |

1. **Question of Interest**

The presented table shows metrics for evaluating various classification methods at the 0.5 threshold level. These findings are important in assessing the accuracy of the models in predicting heart failure. The test error rate, depicting the proportion of misclassified predictions, serves as a crucial indicator. A diminished test error rate signifies heightened accuracy in distinguishing individuals who are either at risk or not at risk of heart failure. The False Positive Rate (FPR) shows instances when the model incorrectly predicts heart failure in the absence of the condition. On the other hand, the False Negative Rate (FNR) highlights instances where the model fails to identify heart failure when it is actually present. The utilization of these three measurements proves highly beneficial when selecting the optimal model for heart failure prediction. Minimizing false positives and false negatives is pivotal. This strategic approach ensures that interventions and resources are judiciously allocated to individuals at the highest risk, emphasizing the significance of precision in predicting heart failure.

1. **Discussion**

Analyzing the model comparisons above yields several key observations. The random forest model stands out with the lowest test error rate (0.24), False Positive Rate (FPR) of 0.1886792, and False Negative Rate (FNR) of 0.3636364 among the three models. These metrics collectively signify that the random forest model performs better in predicting heart failure compared to its counterparts. Conversely, the recursive binary splitting model demonstrates inferior performance among the three. This outcome is likely attributed to the model's limited complexity, highlighting its challenges in capturing the intricacies present in the dataset.

1. **Challenges**

One of the main challenges that we faced in creating these models was establishing a fair threshold when calculating the error rates, false positives, and false negatives. The threshold determines the probability above which an observation is classified as positive (e.g., having heart failure), and below which it is classified as negative. Balancing between the sensitivity and specificity was difficult, as the real-life implications of incorrectly predicting heart failure is very serious.

**References:**

*Anemia and heart failure: Guidance for clinicians and Trialists*. American College of Cardiology. (n.d.). https://www.acc.org/Latest-in-Cardiology/Articles/2021/07/06/12/18/Anemia-and-Heart-Failure#:~:text=Anemia%2C%20the%20reduction%20in%20the,key%20function%20of%20the%20circulation

Bots, S. H., Peters, S. A. E., & Woodward, M. (2017, March 1). *Sex differences in coronary heart disease and stroke mortality: A global assessment of the effect of ageing between 1980 and 2010*. BMJ Global Health. https://gh.bmj.com/content/2/2/e000298

Centers for Disease Control and Prevention. (2022, June 20). *Diabetes and Your Heart*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html#:~:text=Over%20time%2C%20high%20blood%20sugar,and%20can%20damage%20artery%20walls

Chen, X., Jin, H., Wang, D., Liu, J., Qin, Y., Zhang, Y., Zhang, Y., & Xiang, Q. (2023, February 28). *Serum creatinine levels, traditional cardiovascular risk factors and 10-year cardiovascular risk in Chinese patients with hypertension*. Frontiers. https://www.frontiersin.org/articles/10.3389/fendo.2023.1140093/full#:~:text=The%20relationship%20between%20creatinine%20and,risk%20factors%20in%20hypertensive%20patients

U.S. Department of Health and Human Services. (n.d.). *Heart Health and Aging*. National Institute on Aging. https://www.nia.nih.gov/health/heart-health-and-aging

2022 AHA/ACC/HFSA guideline for the management of heart failure ... (n.d.). https://www.ahajournals.org/doi/10.1161/CIR.0000000000001062