Heart Failure Analysis R

Besfort Sadiki & Faruq Gaji

2025-05-08

### Introduction

* This report analyzes clinical records of 299 heart failure patients from Faisalabad Institute of Cardiology (2015) to identify key mortality predictors.
* Using logistic regression and LASSO models, we evaluate 13 variables— including age, ejection fraction, creatinine levels, and follow-up time—to determine their impact on survival outcomes.
* The analysis reveals which factors are statistically significant (e.g., age, kidney function)
* Clinically irrelevant (e.g., diabetes, smoking), offering actionable insights for targeted patient care.

#### **Dataset Overview:**.

* Medical records of 299 heart failure patients collected at Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad, Punjab, Pakistan, during April–December 2015.
* Patients had left ventricular systolic dysfunction. Consisted of 105 women and 194 men, aged between 40 and 95 years old.

#### **Features:** features including clinical, body, and lifestyle information.

**Binary features:**.

* **Anaemia** defined as haematocrit levels lower than 36%.
* **High Blood Pressure** is where the force of blood against artery walls is too high.
* **Diabetes, Sex (i.e, Male or Female), and Smoking.**
* **Creatinine phosphokinase (CPK)** indicates the level of CPK enzyme in blood, possibly indicating heart failure or injury with high levels.
* **Ejection fraction** measures the percentage of blood pumped out by the left ventricle with each contraction.
* **Serum creatinine** indicates kidney function; high levels may suggest renal dysfunction.
* **Serum sodium** test checks sodium levels in the blood, a abnormal levels may indicate heart failure.
* **Death event** feature used as target in binary classification study, indicating if the patient died or survived during the follow-up period (130 days on average).

#### **Problems to be answered**

* What are the major factors regarding the data?
* Why do creatinine and ejection fraction strongly predict mortality, while diabetes and hypertension show no effect?
* How does age quantitatively increase mortality risk?
* Does the male-dominated sample (194 men vs. 105 women) bias the sex-based findings?
* How does cross-validation refine the model compared to standard LASSO?

### Analysis and Result: Log Regression

#### **library**

library(tidyverse)   
 library(ggplot2)   
 library(gamlr)   
 library(dplyr)

##### **Read the Excel File**.

heartfailure <- read.csv("heart\_failure\_clinical\_records\_dataset.csv", strings = T)

##### **Rename the Columns** .

colnames(heartfailure) <- c("Age", "Anaemia", "CPK", "Diabetes",   
 "Ejection Fraction", "High Blood Pressure",   
 "Platelets", "Creatinine", "Sodium", "Sex", "Smoking", "Time", "Death")

##### **Everything is in Correct Format: Numeric, Factor, etc.**

##### **Convert Anaemia (0=no,1=yes) to categorical variable**.

heartfailure$Anaemia <- as.factor(heartfailure$Anaemia)   
  
heartfailure$Diabetes <- as.factor(heartfailure$Diabetes)  
  
heartfailure$`High Blood Pressure` <- as.factor(heartfailure$`High Blood Pressure`)

##### **Converts Sex (1=male,0=female) to categorical variable**.

heartfailure$Sex <- as.factor(heartfailure$Sex)   
  
heartfailure$Smoking <- as.factor(heartfailure$Smoking)   
  
heartfailure$Death <- as.factor(heartfailure$Death)

#### **Fit Log Regression model**.

log\_reg <- glm(Death ~ Age + Anaemia + CPK + Diabetes +   
 `Ejection Fraction` + `High Blood Pressure` + Platelets +   
 Creatinine + Sodium + Sex + Smoking + Time,   
 data = heartfailure, family = binomial)

##### **Coef estimates with standard errors**

coef(log\_reg)

## (Intercept) Age Anaemia1   
## 1.018493e+01 4.741907e-02 -7.470452e-03   
## CPK Diabetes1 `Ejection Fraction`   
## 2.222294e-04 1.451498e-01 -7.666250e-02   
## `High Blood Pressure`1 Platelets Creatinine   
## -1.026794e-01 -1.199624e-06 6.660933e-01   
## Sodium Sex1 Smoking1   
## -6.698107e-02 -5.336580e-01 -1.349222e-02   
## Time   
## -2.104463e-02

**Positive coef (increased mortality risk)**.

* **Age (+0.047):** Older patients have higher mortality risk.
* **Creatinine (+0.666):** Higher creatinine levels result in much higher risk

**Negative coef (decrease mortality risk)**.

* **Ejection Fraction (-0.077):** Better heart functions lowers risk.
* **Time (-0.021):** Longer follow-up result in lower mortality

**Zero or Near-Zero coef (No effect)**.

* **Anaemia (-0.007)**
* **Smoking(-0.013);** these predictors doesn’t significantly affect mortality risk

#### **Model Summary**.

summary(log\_reg)

##   
## Call:  
## glm(formula = Death ~ Age + Anaemia + CPK + Diabetes + `Ejection Fraction` +   
## `High Blood Pressure` + Platelets + Creatinine + Sodium +   
## Sex + Smoking + Time, family = binomial, data = heartfailure)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.018e+01 5.657e+00 1.801 0.071774 .   
## Age 4.742e-02 1.580e-02 3.001 0.002690 \*\*   
## Anaemia1 -7.470e-03 3.605e-01 -0.021 0.983467   
## CPK 2.222e-04 1.779e-04 1.249 0.211684   
## Diabetes1 1.451e-01 3.512e-01 0.413 0.679380   
## `Ejection Fraction` -7.666e-02 1.633e-02 -4.695 2.67e-06 \*\*\*  
## `High Blood Pressure`1 -1.027e-01 3.587e-01 -0.286 0.774688   
## Platelets -1.200e-06 1.889e-06 -0.635 0.525404   
## Creatinine 6.661e-01 1.815e-01 3.670 0.000242 \*\*\*  
## Sodium -6.698e-02 3.974e-02 -1.686 0.091855 .   
## Sex1 -5.337e-01 4.139e-01 -1.289 0.197299   
## Smoking1 -1.349e-02 4.126e-01 -0.033 0.973915   
## Time -2.104e-02 3.014e-03 -6.981 2.92e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 375.35 on 298 degrees of freedom  
## Residual deviance: 219.55 on 286 degrees of freedom  
## AIC: 245.55  
##   
## Number of Fisher Scoring iterations: 6

##### **Significant preditors from the mortality (p < 0.07)**.

* **Age:** (Beta: 0.047, p = 0.003)
* **Ejection Fraction:** (Beta:-0.076 , p = 0.00000267)
* **Creatinine:** (Beta: 0.666 , p = 0.00024)
* **Time:** (Beta: -0.021, p = 0.0000000000029)

##### **Non-Significant Predictors from the mortality**

* **Anaemia:** (Beta: 0.0075 , P = 0.98 )
* **CPK:** (Beta: 0.000222 , P = 0.21 )
* **Diabetes:** (Beta: 0.0145 , P = 0.68 )
* **High Blood Pressure:** (Beta: -0.0103 , P = 0.77)
* **Platelets:** (Beta: 0.0000012 , P = 0.53 )
* **Smoking Status:** (Beta: -0.013 , P = 0.97)
* **Sex:** (Beta: -0.53 , P = 0.20)

##### **Unexpected Findings**.

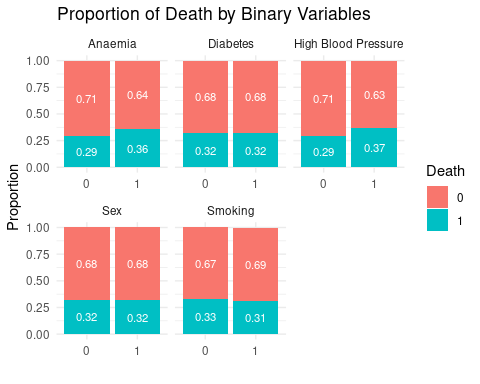
* Diabetes showed no significant effect

##### **Model Fit indicators**

* **Null Deviance (375.35)** on 298 Degrees of Freedom Very High Number: So the model starts with a lot of uncertainty.
* **Residual Deviance (219.55)** on 286 Degrees of Freedom Big drop: Model improved by reducing risk factors
* **AIC** helps you choose the simplest, most accurate model AIC (245.55) Out of 286.

#### **Bar Chart: Binary Variables vs. Death**

heartfailure %>%  
 pivot\_longer(cols = c(Anaemia, Diabetes, `High Blood Pressure`, Sex, Smoking),   
 names\_to = "Variable", values\_to = "Value") %>%  
 group\_by(Variable, Value, Death) %>%  
 summarise(Count = n(), .groups = "drop") %>%  
 group\_by(Variable, Value) %>%  
 mutate(Proportion = Count / sum(Count)) %>%  
 ggplot(aes(x = factor(Value), y = Proportion, fill = factor(Death))) +  
 geom\_bar(stat = "identity", position = "fill") +  
 geom\_text(aes(label = round(Proportion, 2)),  
 position = position\_fill(vjust = 0.5),  
 color = "white",  
 size = 3) +  
 facet\_wrap(~ Variable, scales = "free\_x") +  
 labs(title = "Proportion of Death by Binary Variables",   
 y = "Proportion", x = "", fill = "Death") +  
 theme\_minimal()



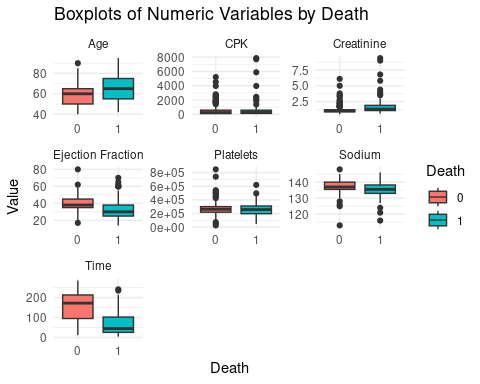
numeric\_vars <- c("Age", "CPK", "Ejection Fraction", "Platelets", "Creatinine", "Sodium", "Time")

#### Inference

* Binary variables **(e.g., diabetes, smoking)** show no clear mortality trend (proportions of death ~50% across categories).
* **Sex** has a slight imbalance (higher male mortality), but statistical tests later confirm it’s insignificant.

#### **Boxplot: Numeric Variables vs. Death**

heartfailure %>%  
 pivot\_longer(cols = all\_of(numeric\_vars), names\_to = "Variable", values\_to = "Value") %>%  
 ggplot(aes(x = Death, y = Value, fill = Death)) +  
 geom\_boxplot() +  
 facet\_wrap(~ Variable, scales = "free") +  
 theme\_minimal() +  
 labs(title = "Boxplots of Numeric Variables by Death")



#### Inference.

* **Age, ejection fraction, creatinine, and time** exhibit stark differences between survival/death groups:
* Increase in Age/Creatinine results in Increase in Mortality (p < 0.001).
* Decrease in **Ejection Fraction/Time** results in Increase in Mortality (p < 0.001).
* CPK, platelets, sodium: Overlapping distributions = weak predictive power.

#### **Analysis and Result: Lasso Model**

##### **Lasso**

lasso\_x <- model.matrix(Death ~ Age + Anaemia + CPK + Diabetes + `Ejection Fraction` + `High Blood Pressure` + Platelets + Creatinine + Sodium + Sex + Smoking + Time, data = heartfailure)[,-1]   
  
lasso\_y <- heartfailure$Death  
  
lasso\_model <- gamlr(lasso\_x, lasso\_y, family = "binomial")

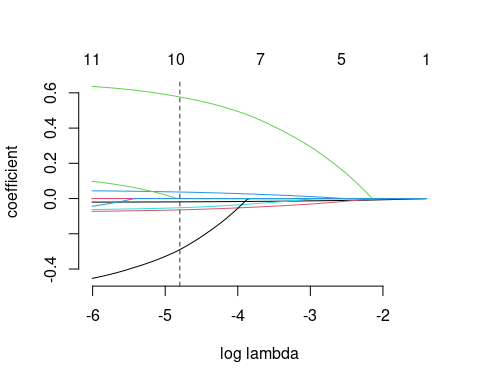
##### **Coefficients**

coef(lasso\_model)

## 13 x 1 sparse Matrix of class "dgCMatrix"  
## seg74  
## intercept 7.943057e+00  
## Age 3.703710e-02  
## Anaemia1 .   
## CPK 1.329507e-04  
## Diabetes1 .   
## `Ejection Fraction` -6.411914e-02  
## `High Blood Pressure`1 .   
## Platelets -3.247326e-07  
## Creatinine 5.765585e-01  
## Sodium -5.221385e-02  
## Sex1 -2.897483e-01  
## Smoking1 .   
## Time -1.876883e-02

##### **Plot**

plot(lasso\_model)



##### **Variable Interpretations (Log\_Odds & Odds Change) from LASSO MODEL**

##### **Assuming all other variables are held constant.**

##### **Age**

* + **Log\_odds:** For every one year increase in age, the log\_odds of death increase by (0.037)
  + **Odds:** The odds of death increase by about 3.8% for each additional year of age. (1.038)

##### **CPK (Creatine Phosphokinase)**

* + **Log\_odds:** For every one unit increase in CPK, the log-odds of death increase by (0.00013)
  + **Odds:** The odds of death increase by about 0.013% per unit increase in CPK. (1.00013)

##### **Ejection Fraction**

* + **Log\_odds:** For every one unit increase in ejection fraction, the log\_odds of death decrease by (0.064)
  + **Odds:** The odds of death decrease by about (6.2%) per unit increase. (0.938)

##### **Platelets**

* + **Log\_odds:** For every one unit increase in platelet count, the log\_odds of death decrease by (0.00000032)
  + **Odds:**The odds of death decrease by about (0.000032%) per unit increase in platelets. (0.99999968)

##### **Creatinine**

* + **Log\_odds:** For every one unit increase in creatinine, the log\_odds of death increase by (0.577)
  + **Odds:** 88 The odds of death increase by about 78% per unit increase in creatinine (1.78)

##### **Sodium**

* + **Log\_odds:** For every one unit increase in sodium, the log\_odds of death decrease by (0.052)
  + **Odds:** The odds of death decrease by about (5.1%) per unit increase. (0.949)

##### **Sex (1 = male)**

* + **Log\_odds:** Being male decreases the log\_odds of death by (0.290) compared to being female
  + **Odds:** Males have about (25.2%) lower odds of death than females. (0.748)
  + May need deeper clinical context; could reflect sample specific patterns.

##### **Time (Follow-up Days)**

* + **Log\_odds:** For every additional day of follow up, the log\_odds of death decrease by (0.0188)
  + **Odds:** The odds of death decrease by about 1.9% for each additional day survived. (0.981)

#### Notes

##### These had no added predictive value at the selected regularization level:

* + **Anaemia, Diabetes, High Blood Pressure, and Smoking.** Lasso decided these were not contributing enough once other variables were accounted for, so they were excluded.
  + **Death** is more likely with older age, higher creatinine, lower ejection fraction, lower sodium, possibly higher CPK.
  + **Sodium:** Lowers pump and blood flow going to the heart. Not likely to lead to heart failure

#### **Analysis and Result: Lasso vs. Cross-Validated Lasso**

##### **Lasso vs. Cross-validated Lasso**

* **Lasso model** identified key predictors of death, which highlighted **Age, CPK, Ejection Fraction, Platelets, Creatinine, Sodium, Sex, and Time** as significant variables.
* **Cross-Validated Lasso** Is more rigorously and selects variables by evaluating prediction performance on unseen data. the model retained only **Age, Ejection Fraction, Creatinine, and Time.** Each with smaller effect sizes.

#### **Cross Validation Model**.

cv\_model <- cv.gamlr(lasso\_x, lasso\_y, family = "binomial")

#### **Coefficients**

coef(cv\_model)

## 13 x 1 sparse Matrix of class "dgCMatrix"  
## seg31  
## intercept 0.856310787  
## Age 0.006887374  
## Anaemia1 .   
## CPK .   
## Diabetes1 .   
## `Ejection Fraction` -0.025076472  
## `High Blood Pressure`1 .   
## Platelets .   
## Creatinine 0.236390013  
## Sodium .   
## Sex1 .   
## Smoking1 .   
## Time -0.012122028

#### **Interpreting Cross Validation Coefficients**

##### **Positive Coefficients**.

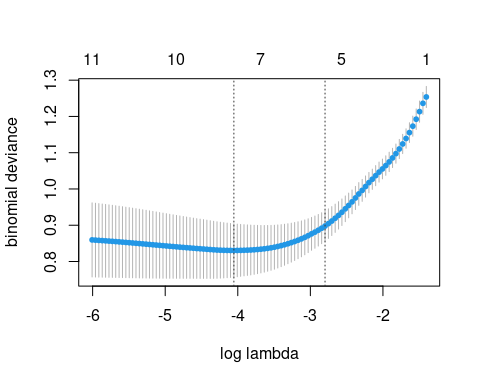
* **Age (0.0148):** Each additional year of age increases log-odds of death by 0.0148.
* **Creatinine (0.3358):** Higher creatinine levels strongly increase mortality risk.

##### **Negative Coefficients**

* **Ejection Fraction (-0.0356):** Better heart function reduces mortality risk.
* **Time (-0.0138):** Longer follow-up duration correlates with improved survival.
* **Sodium (-0.00945):** Higher serum sodium levels may indicate better fluid/electrolyte balance.

#### **Plot**

plot(cv\_model)



#### **Inferences:**

* Each colored line represents a variable in our model.
* On the x-axis, we have log(lambda); the regularization strength.
* As lambda increases (moving left), the model shrinks coefficients toward zero.
* The vertical dashed line is the lambda value selected by cross-validation, this is the point that offers the best predictive performance without over-fitting.
* To the right of that line (smaller lambda), more variables are included with non-zero coefficients.
* As we move closer to the left, only the most important predictors survive- regularization, others are shrunk exactly to zero and dropped.

##### **Suggestion**

* **CPK, Platelets, Sodium, and Sex,**
  + Are insignificant and were removed during cross-validation.
  + they were not essential to accurately distinguishing between survival and death outcomes in this dataset.

#### **Dimensionality Reduction (PCA & Clustering)**

##### **PCA on Numeric Variables**

heartfailure2 <- heartfailure %>%  
 select(Age, CPK, `Ejection Fraction`, Platelets, Creatinine, Sodium, Time)

| Conduct PCA on the LASSO given Variables |  |
| --- | --- |
| PC1 | Age |
| PC2 | CPK |
| PC3 | Ejection Fraction |
| PC4 | Platelets |
| PC5 | Creatinine |
| PC6 | Sodium |
| PC7 | Time |

scaled\_data <- scale(heartfailure2)

#### Perform PCA on numeric variables

pca\_result <- prcomp(scaled\_data, center = TRUE, scale. = TRUE)

#### Summary of PCA

summary(pca\_result)

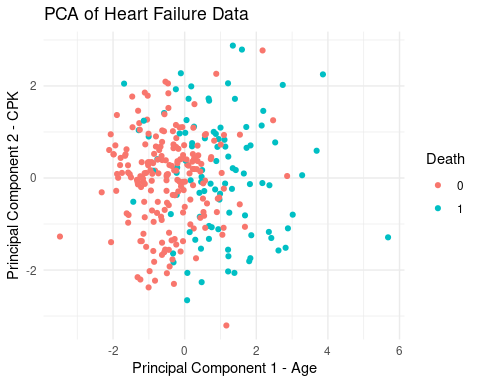
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.2143 1.0842 1.0146 0.9830 0.9422 0.8587 0.8538  
## Proportion of Variance 0.2107 0.1679 0.1471 0.1380 0.1268 0.1053 0.1041  
## Cumulative Proportion 0.2107 0.3786 0.5257 0.6637 0.7905 0.8959 1.0000

#### Visualize the first two principal components

pca\_data <- as.data.frame(pca\_result$x)  
pca\_data$Death <- heartfailure$Death #Add Death for visualization

#### Visualization of Age and CPK

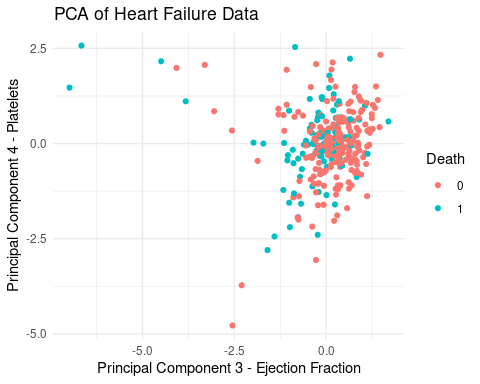
ggplot(pca\_data, aes(x = PC1, y = PC2, color = Death)) +  
 geom\_point() +  
 labs(title = "PCA of Heart Failure Data", x = "Principal Component 1 - Age", y = "Principal Component 2 - CPK") +  
 theme\_minimal()



* Smaller Age and Lower CPK Red dots, Less chance of death.
* Higher age and Higher CPK blue dots, More chance of death.
* Very visible that these are the most significant factors and ranked PCA 1 and PCA 2.

#### Visulization of Ejection Fraction and Platelets

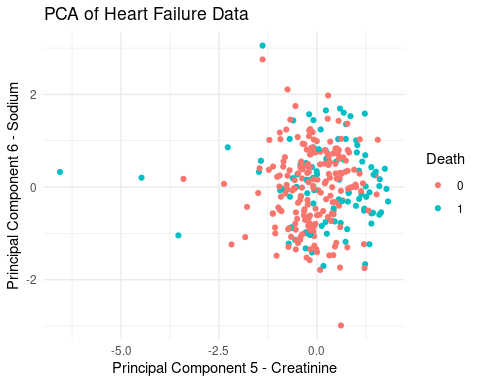
ggplot(pca\_data, aes(x = PC3, y = PC4, color = Death)) +  
 geom\_point() +  
 labs(title = "PCA of Heart Failure Data", x = "Principal Component 3 - Ejection Fraction", y = "Principal Component 4 - Platelets") +  
 theme\_minimal()



* Mix; difficult to tell however lower Ejection Fraction Visible

#### Visulization of Creatinine and Sodium

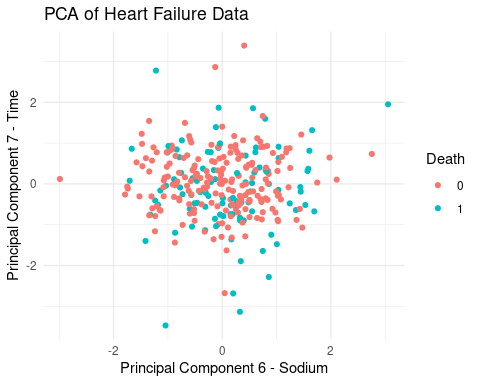
ggplot(pca\_data, aes(x = PC5, y = PC6, color = Death)) +  
 geom\_point() +  
 labs(title = "PCA of Heart Failure Data", x = "Principal Component 5 - Creatinine", y = "Principal Component 6 - Sodium") +  
 theme\_minimal()



* Mix; both set around 0.

#### Visulization of Sodium and Time

ggplot(pca\_data, aes(x = PC6, y = PC7, color = Death)) +  
 geom\_point() +  
 labs(title = "PCA of Heart Failure Data", x = "Principal Component 6 - Sodium", y = "Principal Component 7 - Time") +  
 theme\_minimal()

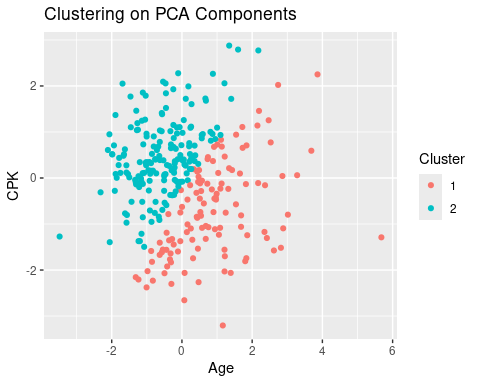


* Mix, both set around 0.

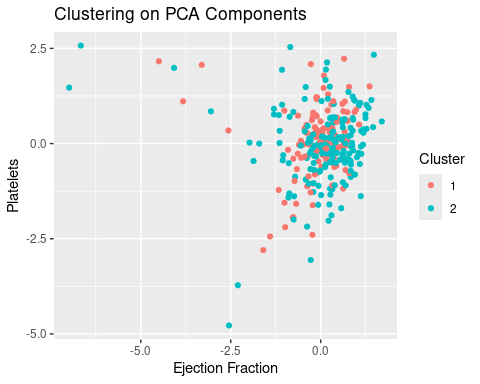
#### K-means clustering on PCA’s

set.seed(123)  
kmeans\_result <- kmeans(pca\_data[,1:4], centers = 2)  
  
pca\_data$Cluster <- as.factor(kmeans\_result$cluster)

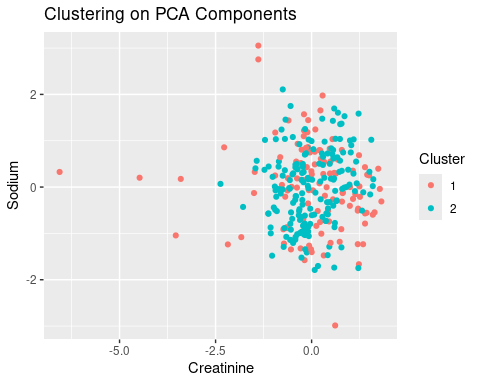
ggplot(pca\_data, aes(x = PC1, y = PC2, color = Cluster)) +  
 geom\_point() +  
 labs(title = "Clustering on PCA Components", x = "Age", y = "CPK")



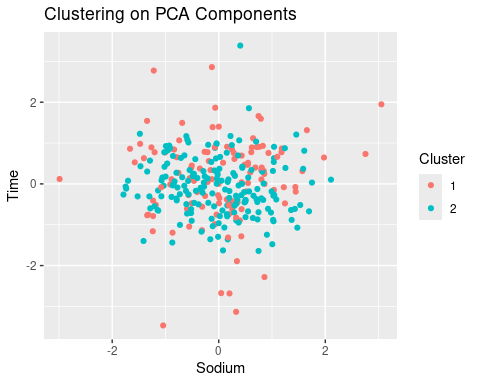
ggplot(pca\_data, aes(x = PC3, y = PC4, color = Cluster)) +  
 geom\_point() +  
 labs(title = "Clustering on PCA Components", x = "Ejection Fraction", y = "Platelets")



ggplot(pca\_data, aes(x = PC5, y = PC6, color = Cluster)) +  
 geom\_point() +  
 labs(title = "Clustering on PCA Components", x = "Creatinine", y = "Sodium")



ggplot(pca\_data, aes(x = PC6, y = PC7, color = Cluster)) +  
 geom\_point() +  
 labs(title = "Clustering on PCA Components", x = "Sodium", y = "Time")



* Clustering Visuals show minor changes for the classified data points.

#### **Conclusion**

* In Conclusion, after running a **logistical regression** from our heartfailure data, it resulted in Age, Ejection Fraction, Creatinine, & Time being the significant factors.
* However, after running a **LASSO model** to initially validate the data, the result was Age, Anaemial, CPK, Ejection Fraction, Platelets, Creatinine, Sodium, Sex, and Time were significant variables to Death.
* Furthermore, to validate the data’s relationship, we created a **Cross Validation** that showed only Age, Ejection Fraction, Creatinine, & Time to be the ultimate significant variables in correlation to Death.
* Lastly, a **Principle Component Analysis (PCA)** and **Clustering** was run to show the ranked variance among the key variables within the data.