Project_CS1380_PN

Phuong Nguyen

2023-05-24

Load + Clean data

The data set “heart_failure” features 299 patients with their conditions and status related to Heart Failure in 2005.

```
heart <- read.csv("heart_failure.csv", header = TRUE)
head(heart)
```

```
##   age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1  75       0                      582        0                20
## 2  55       0                     7861        0                38
## 3  65       0                      146        0                20
## 4  50       1                      111        0                20
## 5  65       1                      160        1                20
## 6  90       1                       47        0                40
##   high_blood_pressure platelets serum_creatinine serum_sodium sex smoking time
## 1                   1    265000              1.9          130   1       0    4
## 2                   0    263358              1.1          136   1       0    6
## 3                   0    162000              1.3          129   1       1    7
## 4                   0    210000              1.9          137   1       0    7
## 5                   0    327000              2.7          116   0       0    8
## 6                   1    204000              2.1          132   1       1    8
##   DEATH_EVENT
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

```
ind <- which(is.na(heart))
ind
```

```
## integer(0)
```

```
heart$sex <- factor(heart$sex, levels = c(0,1), labels = c("Male","Female"))
heart$DEATH_EVENT <- factor(heart$DEATH_EVENT, levels = c(0,1), labels = c("No","Yes"))
heart$high_blood_pressure <- factor(heart$high_blood_pressure, levels = c(0,1), labels = c("No H
igh Blood Pressure","High Blood Pressure"))
heart$diabetes <- factor(heart$diabetes, levels = c(0,1), labels = c("No Diabetes", "Diabetes"))
```

The objective of this project is to identify a pattern among patients suffering from Heart Failure, and to eventually develop a predictive model for mortality by using specific underlying factors.

Extract relevant data + visualization

- Examine: Age, Have high blood pressure, Have diabetes, Level of creatinine_phosphokinase (indicator of heart failure), Ejection fraction (percentage of blood leaving the heart per contraction), Gender and Mortality
- For Age: Usually heart-related illnesses happen more frequently in elderly age group. However, in recent studies, it is suggested that the age range for having heart conditions is shifting “younger”. From the original value, if the age is under 45, it is considered “young”; between 45 and 65 is “middle-age”, and over 65 is consider “senior”.

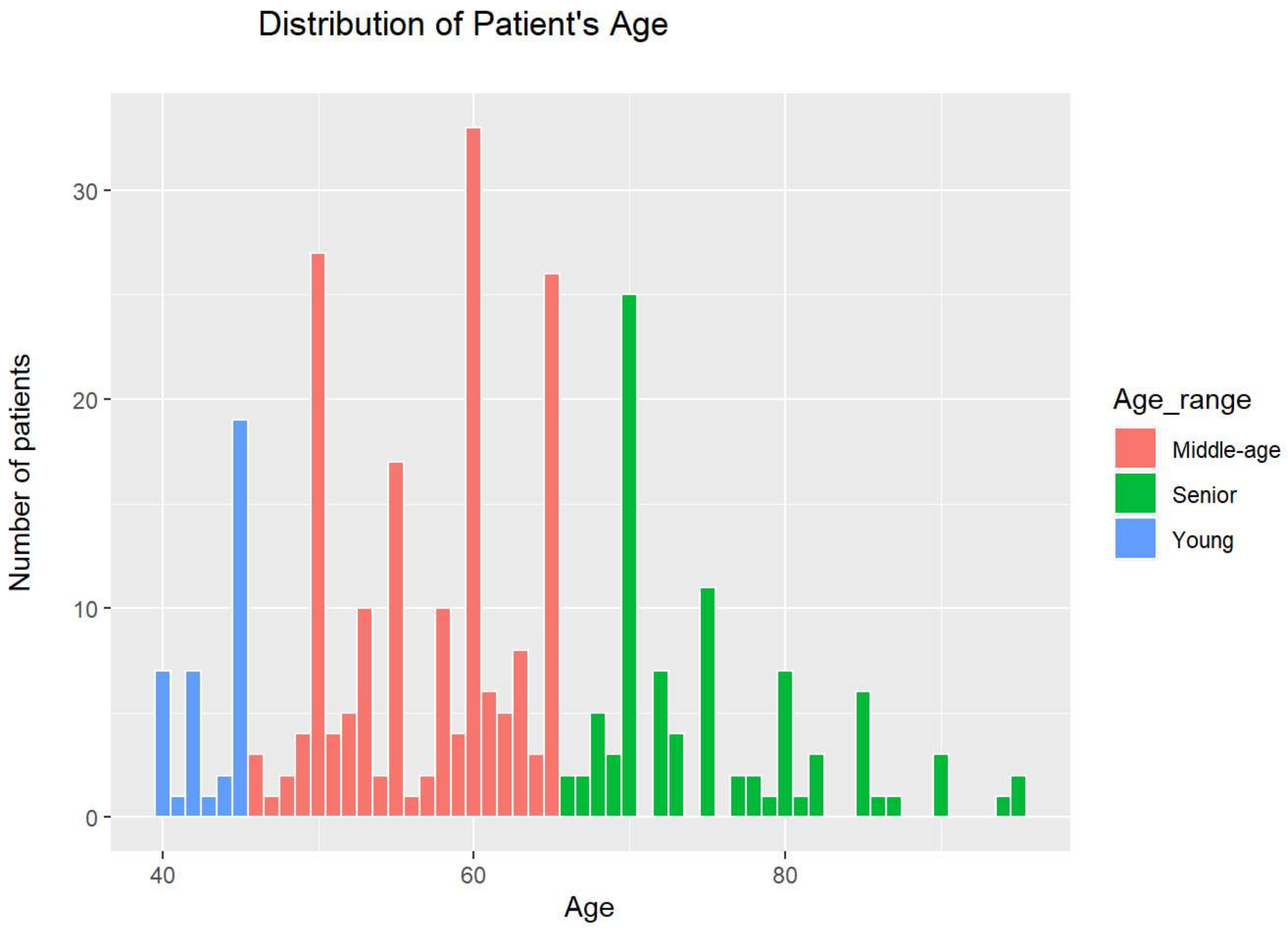
```
library(tidyverse)
library(dplyr)
new_heart <- heart %>%
  select(c(age,high_blood_pressure,diabetes,creatinine_phosphokinase, ejection_fraction, sex, DE
ATH_EVENT)) %>%
  mutate(Age_range = case_when(age <= 45 ~ "Young",
                               age <= 65 ~ "Middle-age",
                               age > 65 ~ "Senior")) %>%
  rename(CPK_level = creatinine_phosphokinase, HBP = high_blood_pressure, Ejection_percentage =
ejection_fraction, Diabetes = diabetes, Mortality = DEATH_EVENT, Gender = sex, Age = age)
head(new_heart)
```

```
##   Age                    HBP    Diabetes CPK_level Ejection_percentage Gender
## 1  75    High Blood Pressure No Diabetes       582                  20 Female
## 2  55 No High Blood Pressure No Diabetes      7861                  38 Female
## 3  65 No High Blood Pressure No Diabetes       146                  20 Female
## 4  50 No High Blood Pressure No Diabetes       111                  20 Female
## 5  65 No High Blood Pressure    Diabetes       160                  20   Male
## 6  90    High Blood Pressure No Diabetes        47                  40 Female
##   Mortality  Age_range
## 1       Yes     Senior
## 2       Yes Middle-age
## 3       Yes Middle-age
## 4       Yes Middle-age
## 5       Yes Middle-age
## 6       Yes     Senior
```

Distribution of patients’ Age in the study

We start by examining the distribution of patients’ age in this data set

```
ggplot(new_heart, aes(x = Age, fill = Age_range)) +
  geom_histogram(binwidth = 1, colour = "white") +
  ylab("Number of patients\n") +
              ggtitle("               Distribution of Patient's Age\n")
```

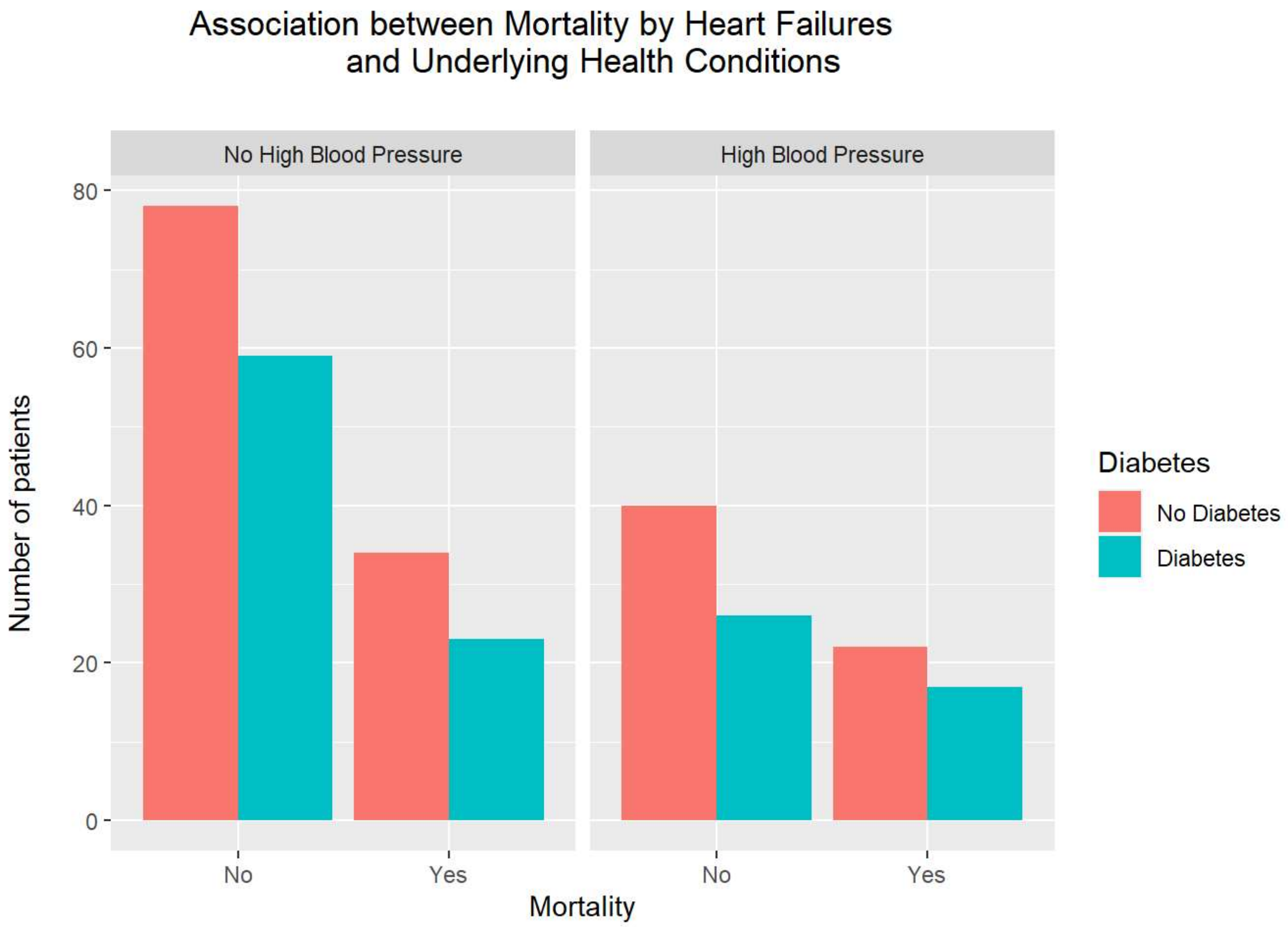


The histogram is right-skewed. The majority of patients with heart failures are middle-age and senior.

Visualise number of paients who have high blood pressure or diabetes and the association to Mortality

Next, underlying health conditions are compared against mortality to see if there is an association.

```
ggplot(new_heart, aes(x = Mortality, fill = Diabetes)) +
  geom_bar(position = "dodge") +
  facet_grid(.~ HBP) + ylab("Number of patients\n") +
  ggtitle("        Association between Mortality by Heart Failures
                        and Underlying Health Conditions\n")
```

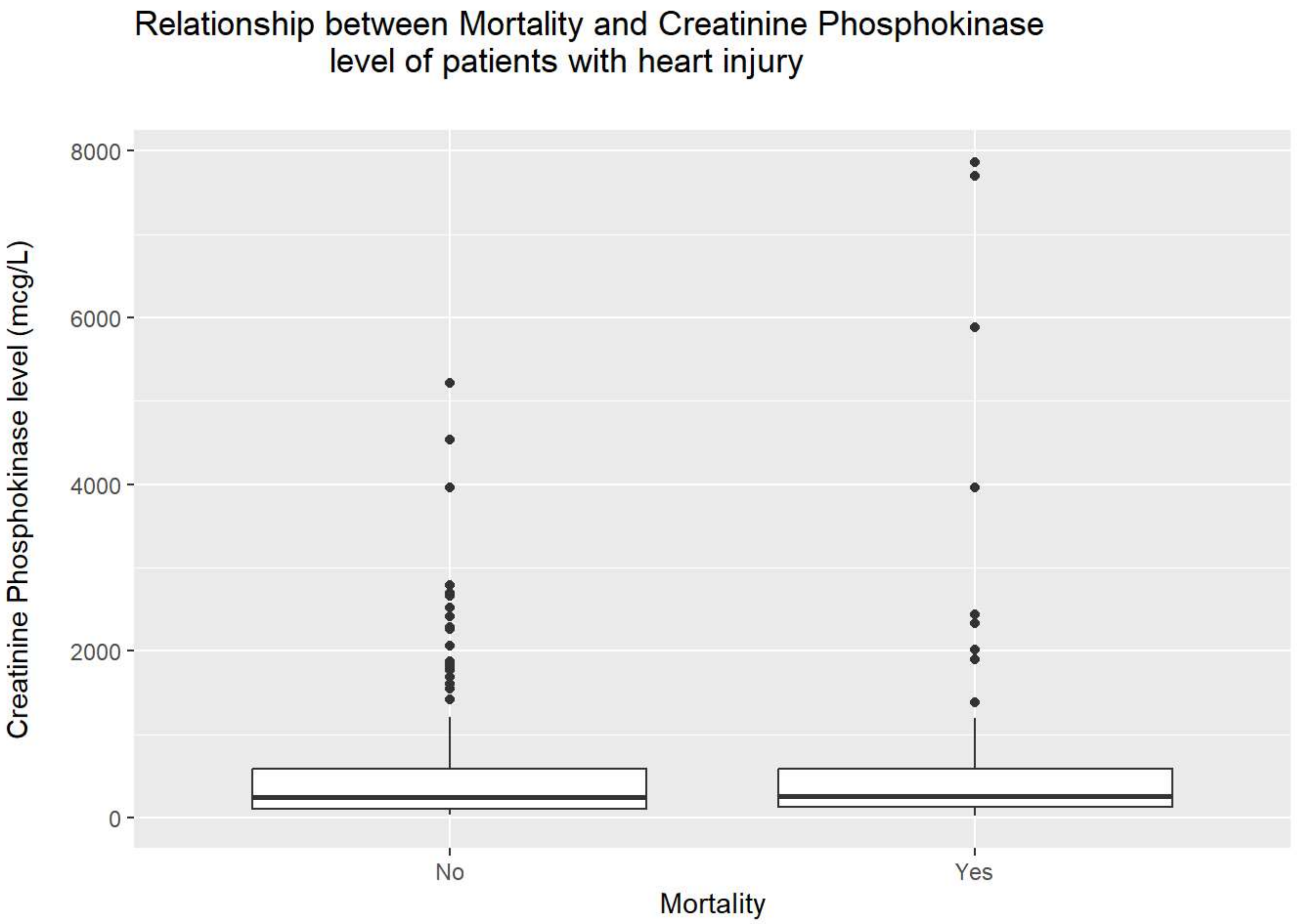


In general, from the bar plot, having diabetes or high-blood-pressure do not contribute significantly to a higher chance of mortality.

Visualize the relationship between Mortality and CPK level in blood

Creatinine phosphokinase (CPK) is an enzyme that is found mainly in muscle tissues. The normal level of CPK in blood usually ranges from 10 -120 mcg/L, and when there is a muscle injury or inflammation, CPK is released causing a surge. Here, the relationship between CPK_level and Mortality is visualized to see if there is a positive association.

```
library(ggplot2)
ggplot(new_heart, aes(x = Mortality, y = CPK_level)) +
  geom_boxplot() +
  ggtitle("Relationship between Mortality and Creatinine Phosphokinase
                  level of patients with heart injury\n") +
  ylab("Creatinine Phosphokinase level (mcg/L)\n")
```

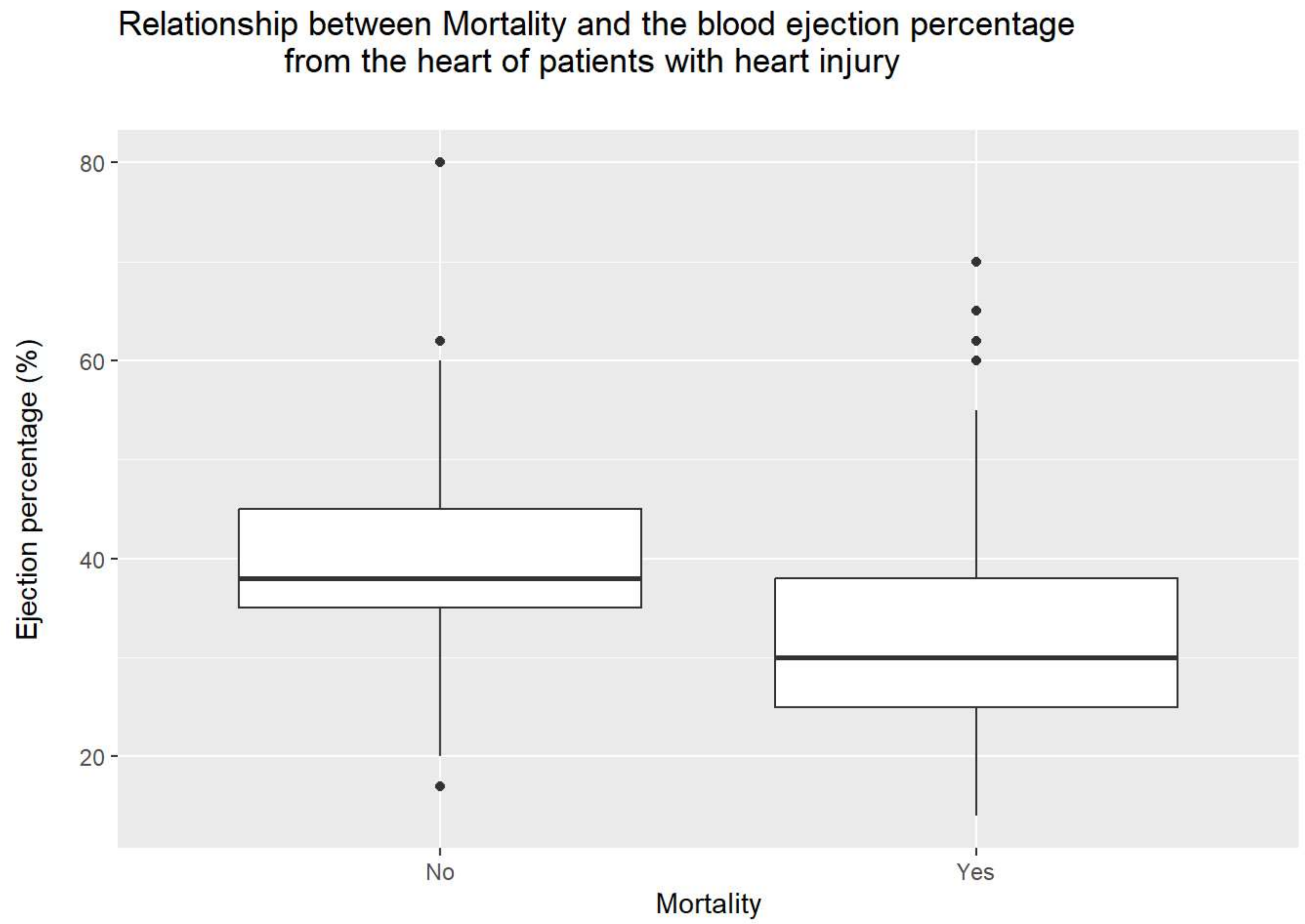


The boxplot shows that in general, there is little difference in the mean of CPK level in blood between patients who survive or die until the next health check-up. A possible explanation for the insignificant difference is because all patients in this study have suffered from heart failures in varied degrees, the CPK values are high already for all patients. There are, however, many outliers of CPK level in the upper region for both groups.

Visualize the relationship between Mortality and Percentage of blood leaving the heart per contraction (Ejection_percentage/EP)

Ejection percentage (EP) is the percentage of blood leaving the heart per contraction. A normal ejection percentage is between 55% and 75%; the range 40 - 55% is considered low function, and under 40% means there is a risk for possible heart failure. The values of Ejection_percentage in this data set are visualized and compared to Mortality to see if there is any relationship.

```
library(ggplot2)
ggplot(new_heart, aes(x = Mortality, y = Ejection_percentage)) +
  geom_boxplot() +
  ggtitle("Relationship between Mortality and the blood ejection percentage
               from the heart of patients with heart injury\n") +
  ylab("Ejection percentage (%)\n")
```



The boxplot shows that between survived and non-survived patients, the average percentage of blood leaving the heart per contraction is lower than that in survived patients. It suggests that with less blood flow from the heart, the risk of mortality is higher. In the group of non-survived patients, there are some outliers in the 60% and above range. This result agrees with medical knowledge that too high of ejection percentage is caused by some issues such as hypertrophic cardiomyopathy, which leads to sudden cardiac arrest.

Summarize of Average CPK, Average EP and Mortality rate by Gender and Age_group

Different genders and Age groups can have an effect to a certain extent on mortality.

```
new_heart %>%
  group_by(Gender, Age_range) %>%
  summarise(Average_CPK = mean(CPK_level), Average_EP = mean(Ejection_percentage), Mortality_Rat
e = mean(Mortality == "Yes"))
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the
## `.groups` argument.
## # A tibble: 6 × 5
## # Groups:   Gender [2]
##   Gender Age_range  Average_CPK Average_EP Mortality_Rate
##   <fct>  <chr>            <dbl>      <dbl>          <dbl>
## 1 Male   Middle-age        500.       38.4         0.333
## 2 Male   Senior            369.       42.2         0.407
## 3 Male   Young             593.       47.8         0.0833
## 4 Female Middle-age        585.       37.4         0.234
## 5 Female Senior            567.       37.1         0.5
## 6 Female Young            1048.       33.5         0.24
```

- From the summary table for Mortality rate among genders, age ranges, average CPK levels and average Ejection percentage, in this study, Senior age group has the highest mortality rate for both males and females, followed by middle-age group and young group.
- However, for females, the mortality rates are equal between the young and middle-age group, with the young group having a significantly higher average CPK level than the middle-age.
- Between males and females, ejection percentage, on average, is higher in males than in females.

Prediction model

There are several types of the CPK enzyme that are associated with different organs. When a person undergoes a heart attack or heart injury, the CPK-2 level is elevated in the blood. Hence, CPK level in the blood can be used as an indicator for heart attack. Additionally, Ejection percentage (EP) is believed to be a contributor to the mortality of patients in heart failure. This project aims to use CPK and EP to predict the mortality rate. The predicted model will then be compared with actual data by using a confusion matrix to see how much accuracy it can achieved.

```
#model 1
pred.model <- glm(Mortality ~ CPK_level + Ejection_percentage, new_heart,
                  family = "binomial")
summary(pred.model)
```

```
##
## Call:
## glm(formula = Mortality ~ CPK_level + Ejection_percentage, family = "binomial",
##     data = new_heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3105  -0.8958  -0.7076   1.1383   2.3263
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.2358019  0.4705135   2.626  0.00863 **
## CPK_level            0.0001138  0.0001255   0.907  0.36457
## Ejection_percentage -0.0560160  0.0126073  -4.443 8.86e-06 ***
## ---
## 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: 351.15  on 296  degrees of freedom
## AIC: 357.15
##
## Number of Fisher Scoring iterations: 4
```

```
#model 2
pred.cpk <- glm(Mortality ~ CPK_level, data = new_heart, family = "binomial")
summary(pred.cpk)
```

```
##
## Call:
## glm(formula = Mortality ~ CPK_level, family = "binomial", data = new_heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1141  -0.8792  -0.8573   1.5084   1.5411
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.8265731  0.1447064  -5.712 1.12e-08 ***
## CPK_level    0.0001297  0.0001218   1.065    0.287
## ---
## 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: 374.23  on 297  degrees of freedom
## AIC: 378.23
##
## Number of Fisher Scoring iterations: 4
```

```
#model 3
pred.ep <- glm(Mortality ~ Ejection_percentage, data = new_heart, family = "binomial")
summary(pred.ep)
```

```
##
## Call:
## glm(formula = Mortality ~ Ejection_percentage, family = "binomial",
##     data = new_heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3320  -0.9146  -0.7201   1.2173   2.3205
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.31169    0.46278   2.834  0.00459 **
## Ejection_percentage -0.05620    0.01258  -4.468 7.88e-06 ***
## ---
## 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: 351.97  on 297  degrees of freedom
## AIC: 355.97
##
## Number of Fisher Scoring iterations: 4
```

Analysis of prediction models 1 & 3

- Three logistic regression model have a common result that the p-value of ejection percentage is statistically significant (<0.001), while the p-value of CPK-level in the all models is statistically insignificant (>0.1). Hence, model 3 and possibly model 1 should be used for prediction of mortality.
- In both model 1 and model 3: When ‘Ejection_percentage’ increases by 1 unit, the log-odd of patients’ mortality decreases by ~0.056 unit
- Below is the prediction of mortality when using model 1 and 3 and a standard cutoff value at 0.5

```
#Model 1
pred_mort_1 <- predict(pred.model, new_heart, type = "response")
pred_lab_1 = if_else(pred_mort_1 > 0.5, "Yes","No")
pred_lab_1 = factor(pred_lab_1, labels = c("No","Yes"))
cfs.mtx.1 = table(pred_lab_1, new_heart$Mortality)
cfs.mtx.1
```

```
##
## pred_lab_1  No Yes
##        No  199  72
##        Yes   4  24
```

```
accuracy.1 = (sum(diag(cfs.mtx.1)))/sum(cfs.mtx.1)
accuracy.1
```

```
## [1] 0.7458194
```

```
sensitivity.1 = cfs.mtx.1[2,2]/sum(cfs.mtx.1[,2])
sensitivity.1
```

```
## [1] 0.25
```

```
specificity.1 = cfs.mtx.1[1,1]/sum(cfs.mtx.1[,1])
specificity.1
```

```
## [1] 0.9802956
```

```
#Model 3
pred_mort_3 <- predict(pred.ep, new_heart, type = "response")
pred_lab_3 = if_else(pred_mort_3 > 0.5, "Yes","No")
pred_lab_3 = factor(pred_lab_3, labels = c("No","Yes"))
cfs.mtx.3 = table(pred_lab_3, new_heart$Mortality)
cfs.mtx.3
```

```
##
## pred_lab_3  No Yes
##        No  200  76
##        Yes   3  20
```

```
accuracy.3 = (sum(diag(cfs.mtx.3)))/sum(cfs.mtx.3)
accuracy.3
```

```
## [1] 0.735786
```

```
sensitivity.3 = cfs.mtx.3[2,2]/sum(cfs.mtx.3[,2])
sensitivity.3
```

```
## [1] 0.2083333
```

```
specificity.3 = cfs.mtx.3[1,1]/sum(cfs.mtx.3[,1])
specificity.3
```

```
## [1] 0.9852217
```

Model 1 has an accuracy of 0.75, a sensitivity of 0.25 and a specificity of 0.98. Model 3 has an accuracy of 0.74, a sensitivity of 0.21 and a specificity of 0.99. A high specificity means the models correctly identify the true negatives, aka the number of survived patients (no mortality). However, the two models have low sensitivity, which means they cannot correctly predict the true positives, aka the number of non-survived patients (mortality).

Examine accuracy by choosing cutoff values

A high cut-off value is prone to produce more false negative (less mortality than actual cases), while a low cut-off value is prone to produce more false positive (more mortality than actual cases). A range of cutoff value is chosen to evaluate the accuracy of model 1 and 3

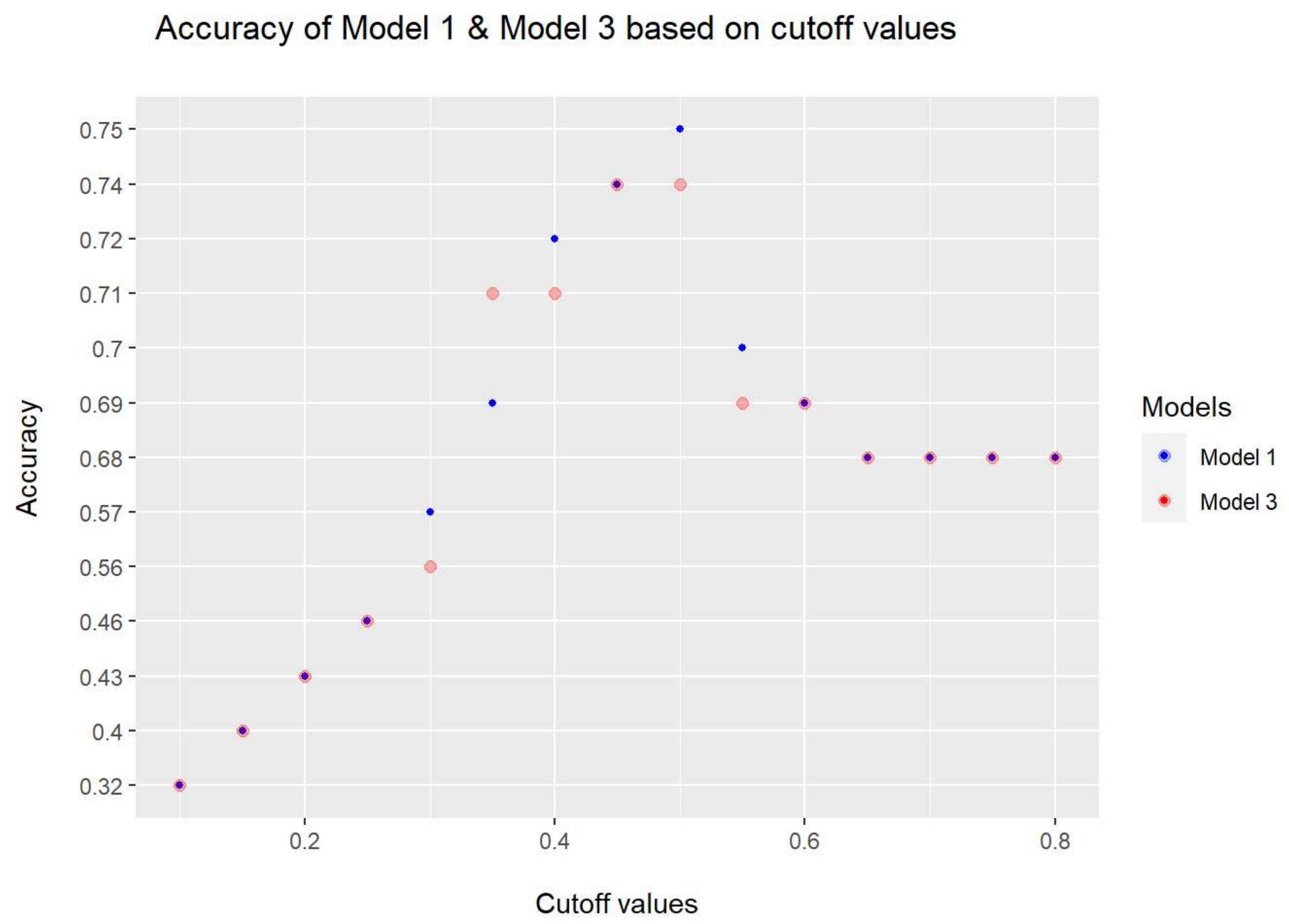
```r
cutoff_seq = seq(from = 0.1, to = 0.8, by = 0.05)

#Model 1
accuracy.1.seq = character(length = length(cutoff_seq))
pred_mort_1 <- predict(pred.model, new_heart, type = "response")
for (i in 1:length(cutoff_seq)){
  cutoff = cutoff_seq[i]
  pred_lab_1 = if_else(pred_mort_1 > cutoff, "Yes","No")
  #pred_lab_1 = factor(pred_lab_1, labels = c("No","Yes"))
  cfs.mtx.1 = table(pred_lab_1, new_heart$Mortality)
  accuracy_1 = (sum(diag(cfs.mtx.1)))/sum(cfs.mtx.1)
  accuracy.1.seq[i] = round(accuracy_1,2)
}

#Model 3
accuracy.3.seq = character(length = length(cutoff_seq))
pred_mort_3 <- predict(pred.ep, new_heart, type = "response")
for (i in 1:length(cutoff_seq)){
  cutoff = cutoff_seq[i]
  pred_lab_3 = if_else(pred_mort_3 > cutoff, "Yes","No")
  #pred_lab_1 = factor(pred_lab_1, labels = c("No","Yes"))
  cfs.mtx.3 = table(pred_lab_3, new_heart$Mortality)
  accuracy_3 = (sum(diag(cfs.mtx.3)))/sum(cfs.mtx.3)
  accuracy.3.seq[i] = round(accuracy_3,2)
}

df = data.frame(Cutoff_Values = cutoff_seq, Accuracy.1 = accuracy.1.seq, Accuracy.3 = accuracy.3.seq)

ggplot(df, aes(x = cutoff_seq)) +
  geom_point(aes(y= accuracy.1.seq, color = "Model 1"), pch = 19, size = 1) +
  geom_point(aes(y=accuracy.3.seq, color = "Model 3"), pch = 19,alpha = 0.3, size = 2) +
  labs(x = "\nCutoff values", y = "Accuracy\n",
       title = "  Accuracy of Model 1 & Model 3 based on cutoff values \n", color = "Models") +
  scale_color_manual(values = c("Model 1" = "blue", "Model 3" = "red"))
```



From the plot of accuracy from two predictive models, it can be observed that the maximum of achievable accuracy is around ~0.75, when the cutoff is at 0.5. Although the accuracy can be classified as "relatively well", more factors should be considered in the logistic regression model or a different type of prediction models should be used to optimize the procedure of heart-injury diagnosis.

References

[1] Larxel. (2020, June 20). Heart failure prediction. Kaggle. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data (https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data)

[2] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5 (https://doi.org/10.1186/s12911-020-1023-5)

[3] Creatine phosphokinase test. Mount Sinai Health System. (n.d.). https://www.mountsinai.org/health-library/tests/creatine-phosphokinase-test# (https://www.mountsinai.org/health-library/tests/creatine-phosphokinase-test#):~:text=Normal%20Results,per%20liter%20(mcg%2FL)

[4] Ejection Fraction: What the Numbers Mean. Pennmedicine.org. (2022, April 13). https://www.pennmedicine.org/updates/blogs/heart-and-vascular-blog/2022/april/ejection-fraction-what-the-numbers-mean# (https://www.pennmedicine.org/updates/blogs/heart-and-vascular-blog/2022/april/ejection-fraction-what-the-numbers-mean#):~:text=Ejection%20fraction%20is%20measured%20as,blood%20and%20may%20be%20failing.