피마 인디언 당뇨병 데이터베이스

Doyoung Kim

2022-08-08

# 당뇨병

#### 혈액 속의 포도당이 세포 속으로 들어가 에너지원으로 이용되지 못하여 혈당이 비정상적으로 올라가는 질환

### 1) 탄수화물 섭취하면 위장에서 소화효소에 의해 탄수화물의 기본 구성성분인 포도당으로 변한 다음 혈액으로 흡수

### 2) 포도당은 기본적인 에너지원. 흡수된 포도당이 세포들에서 이용되기 위해 인슐린 호르몬이 필요

### 3) 인슐린은 췌장의 베타세포에서 분비되어 식사 후 올라간 혈당을 낮추는 기능

### 자료 출처

##### <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

### 변수

##### Pregnancies: 임신 횟수

##### Glucose: 포도당 부하 검사 수치

##### BloodPressure: 혈압(mm Hg)

##### SkinThickness: 팔 삼두근 뒤쪽의 피하지방 측정값(mm)

##### Insulin: 혈청 인슐린(mu U/ml)

##### BMI: 체질량지수(체중(kg)/키(m))^2

##### DiabetesPedigreeFunction: 당뇨 내력 가중치 값

##### Age: 나이

##### Outcome: 클래스 결정 값(0 또는 1)

## 파일 로딩 및 기본 정보 확인

df <- read.csv('./diabetes/diabetes.csv')  
str(df)

## 'data.frame': 768 obs. of 9 variables:  
## $ Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ...  
## $ Glucose : int 148 85 183 89 137 116 78 115 197 125 ...  
## $ BloodPressure : int 72 66 64 66 40 74 50 0 70 96 ...  
## $ SkinThickness : int 35 29 0 23 35 0 32 0 45 0 ...  
## $ Insulin : int 0 0 0 94 168 0 88 0 543 0 ...  
## $ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...  
## $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...  
## $ Age : int 50 31 32 21 33 30 26 29 53 54 ...  
## $ Outcome : int 1 0 1 0 1 0 1 0 1 1 ...

df$Outcome <- factor(df$Outcome)  
str(df)

## 'data.frame': 768 obs. of 9 variables:  
## $ Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ...  
## $ Glucose : int 148 85 183 89 137 116 78 115 197 125 ...  
## $ BloodPressure : int 72 66 64 66 40 74 50 0 70 96 ...  
## $ SkinThickness : int 35 29 0 23 35 0 32 0 45 0 ...  
## $ Insulin : int 0 0 0 94 168 0 88 0 543 0 ...  
## $ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...  
## $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...  
## $ Age : int 50 31 32 21 33 30 26 29 53 54 ...  
## $ Outcome : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 2 ...

## 결측치 확인

sum(is.na(df))

## [1] 0

summary(df)

## Pregnancies Glucose BloodPressure SkinThickness   
## Min. : 0.000 Min. : 0.0 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 1.000 1st Qu.: 99.0 1st Qu.: 62.00 1st Qu.: 0.00   
## Median : 3.000 Median :117.0 Median : 72.00 Median :23.00   
## Mean : 3.845 Mean :120.9 Mean : 69.11 Mean :20.54   
## 3rd Qu.: 6.000 3rd Qu.:140.2 3rd Qu.: 80.00 3rd Qu.:32.00   
## Max. :17.000 Max. :199.0 Max. :122.00 Max. :99.00   
## Insulin BMI DiabetesPedigreeFunction Age   
## Min. : 0.0 Min. : 0.00 Min. :0.0780 Min. :21.00   
## 1st Qu.: 0.0 1st Qu.:27.30 1st Qu.:0.2437 1st Qu.:24.00   
## Median : 30.5 Median :32.00 Median :0.3725 Median :29.00   
## Mean : 79.8 Mean :31.99 Mean :0.4719 Mean :33.24   
## 3rd Qu.:127.2 3rd Qu.:36.60 3rd Qu.:0.6262 3rd Qu.:41.00   
## Max. :846.0 Max. :67.10 Max. :2.4200 Max. :81.00   
## Outcome  
## 0:500   
## 1:268   
##   
##   
##   
##

## 0 <- NA로 바꾸기

df$Glucose[df$Glucose==0] <- NA  
df$BloodPressure[df$BloodPressure==0] <- NA  
df$SkinThickness[df$SkinThickness==0] <- NA  
df$BMI[df$BMI==0] <- NA

## 결측치 그래프

library(VIM)

## 필요한 패키지를 로딩중입니다: colorspace

## 필요한 패키지를 로딩중입니다: grid

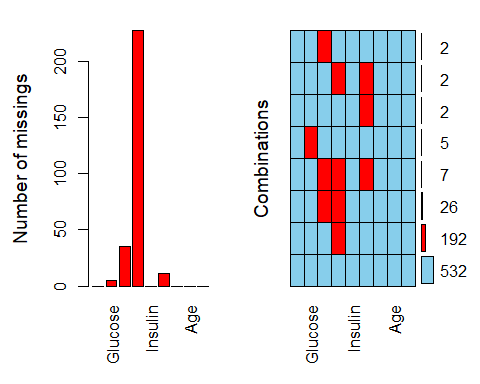
## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##   
## 다음의 패키지를 부착합니다: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

aggr(df, numbers=T, prop=F)

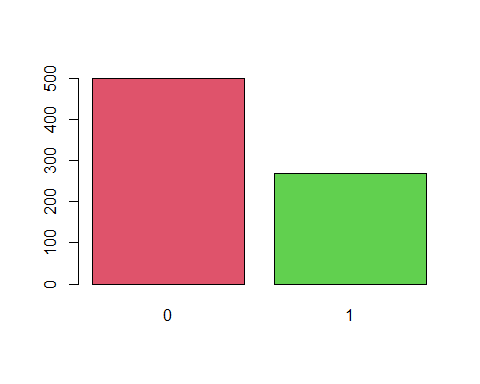


## 정상과 환자 구분하여 NA <- 각 변수 평균으로 채우기

df$Glucose[is.na(df$Glucose) & df$Outcome==0] <- mean(df$Glucose[df$Outcome==0], na.rm=T)  
df$BloodPressure[is.na(df$BloodPressure) & df$Outcome==0] <- mean(df$BloodPressure[df$Outcome==0], na.rm=T)  
df$SkinThickness[is.na(df$SkinThickness) & df$Outcome==0] <- mean(df$SkinThickness[df$Outcome==0], na.rm=T)  
df$BMI[is.na(df$BMI) & df$Outcome==0] <- mean(df$BMI[df$Outcome==0], na.rm=T)  
   
df$Glucose[is.na(df$Glucose) & df$Outcome==1] <- mean(df$Glucose[df$Outcome==1], na.rm=T)  
df$BloodPressure[is.na(df$BloodPressure) & df$Outcome==1] <- mean(df$BloodPressure[df$Outcome==1], na.rm=T)  
df$SkinThickness[is.na(df$SkinThickness) & df$Outcome==1] <- mean(df$SkinThickness[df$Outcome==1], na.rm=T)  
df$BMI[is.na(df$BMI) & df$Outcome==1] <- mean(df$BMI[df$Outcome==1], na.rm=T)

## 정상과 환자 수 그래프

barplot(table(df$Outcome), col = 2:3)



## 정상과 환자 평균 비교

library(psych)  
describe(df[df$Outcome==0, ])[,3]

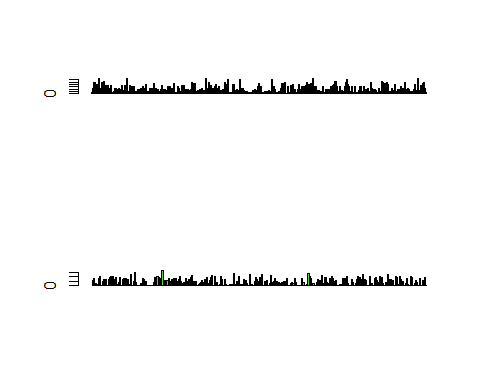
## [1] 3.298000 110.643863 70.877339 27.235457 68.792000 30.859674 0.429734  
## [8] 31.190000 1.000000

describe(df[df$Outcome==1, ])[,3]

## [1] 4.865672 142.319549 75.321429 33.000000 100.335821 35.406767 0.550500  
## [8] 37.067164 2.000000

## 그래프 그리기

par(mfrow=c(2,1))  
barplot(df$Pregnancies[df$Outcome==0], col=2)  
barplot(df$Pregnancies[df$Outcome==1], col=3)

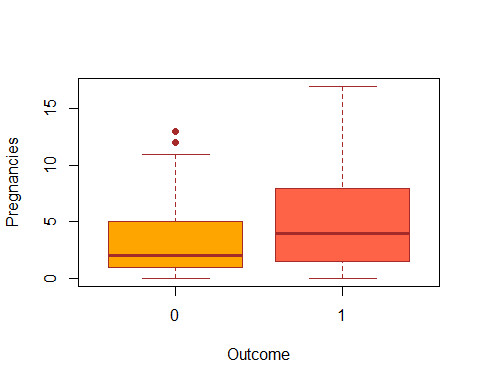


par(mfrow=c(1,1))

## 임신 횟수

##### 임신 중에 분비되는 태반 호르몬이 인슐린의 작용을 방해해 발생되는 인슐린 저항성 때문. 보통은 임신성 당뇨병의 경우 출산을 하게 되면 혈당이 정상화 되지만, 이후에 당뇨병으로 진행할 확률이 정상인의 4배

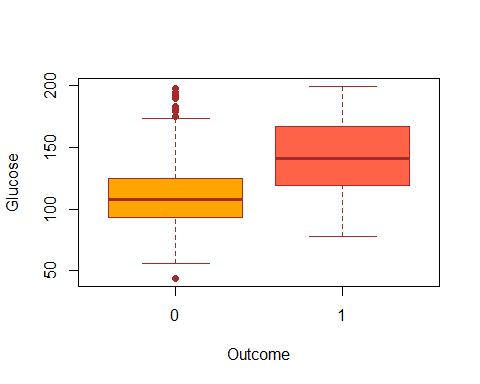
boxplot(Pregnancies ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 포도당 부하 검사 수치

##### 포도당을 부하했을 때, 당뇨병 환자이면 부하된 포도당을 정상인과 같이 신속하게 혈중에서 제거할 수 없다. 140mg/dL 미만 - 정상

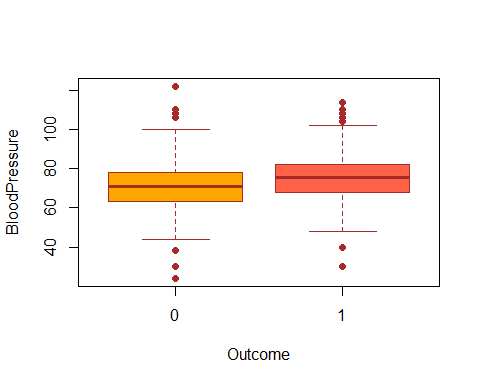
boxplot(Glucose ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 혈압

##### 인슐린에 대한 세포의 저항성이 고혈압이나 당뇨를 일으키는 원인으로 동일하게 작용하기 때문에 고혈압이 있는 사람은 향후 당뇨가 발생할 가능성이 일반인보다 약 2배 높다.

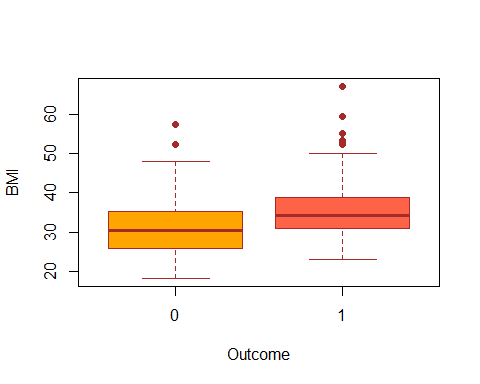
boxplot(BloodPressure ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 체질량지수(체중(kg)/키(m))^2

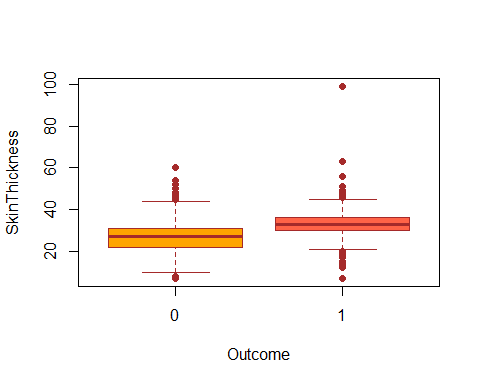
##### 비만으로 인해 혈중 유리지방의 수치가 증가하게 되면서, 근육의 포도당 섭취가 저하하게 되고, 이로 인해 인슐린 감도가 약해지고, 또한 간에서 더 많은 포도당을 생산하게 되며 고혈당증을 더욱 부추기게 된다.

boxplot(BMI ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 팔 삼두근 뒤쪽의 피하지방 측정값(mm)

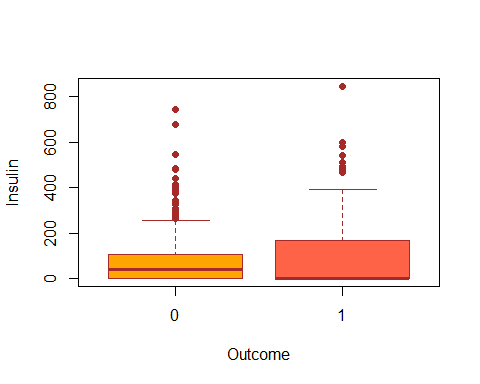
boxplot(SkinThickness ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 혈청 인슐린(mu U/ml)

##### <https://terms.naver.com/entry.naver?docId=500275&cid=60408&categoryId=55558>

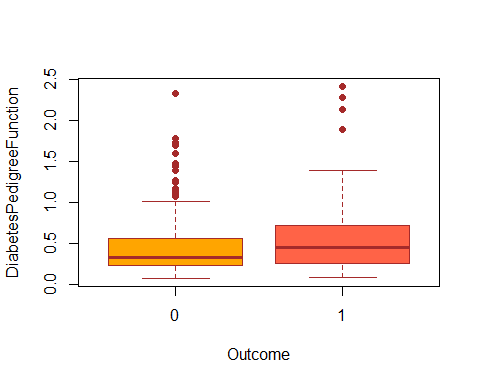
boxplot(Insulin ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 당뇨 내력 가중치 값

##### 부모가 모두 제2형 당뇨병인 경우 자녀에게서 제2형 당뇨병이 발병할 가능성은 30% 정도, 부모 중 한 사람만 제2형 당뇨병인 경우 자녀에게 제2형 당뇨병이 발병할 가능성은 15% 정도다.

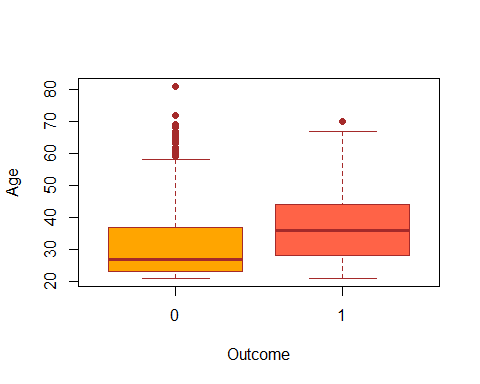
boxplot(DiabetesPedigreeFunction ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## Age: 나이

##### 제1형 당뇨병은 15세 미만의 소아청소년에게 주로 발생하는 당뇨병, 제2형 당뇨병은 30대 이상 주로 발생하는 당뇨병

boxplot(Age ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 환자 수 만큼 정상인 수 샘플링하여 평균 값 비교하기

library(psych)  
set.seed(2022)  
rows <- as.numeric(rownames(df[df$Outcome==0,]))  
df.origin <- data.frame(Outcome=df[rows,]$Outcome,  
 Pregnancies=df[rows,]$Pregnancies,  
 Glucose=df[rows,]$Glucose,  
 BloodPressure=df[rows,]$BloodPressure,  
 SkinThickness=df[rows,]$SkinThickness,  
 Insulin=df[rows,]$Insulin,  
 BMI=df[rows,]$BMI, DiabetesPedigreeFunction=df[rows,]$DiabetesPedigreeFunction,  
 Age=df[rows,]$Age)  
  
idx <- sample(rows, size=268)  
df.sample <- df[idx,]  
  
dfall <- merge(x = df.sample, y = df[df$Outcome==1,], all = T)  
df0 <- dfall[dfall$Outcome==0, ]  
df1 <- dfall[dfall$Outcome==1, ]  
  
data.frame(name=names(df),mean0=describe(df0)[,3],mean1=describe(df1)[,3], max=describe(dfall)[,9])

## name mean0 mean1 max  
## 1 Pregnancies 3.2537313 4.865672 17.00  
## 2 Glucose 111.2262831 142.319549 199.00  
## 3 BloodPressure 71.6455922 75.321429 122.00  
## 4 SkinThickness 27.9317815 33.000000 99.00  
## 5 Insulin 63.0485075 100.335821 846.00  
## 6 BMI 31.2466357 35.406767 67.10  
## 7 DiabetesPedigreeFunction 0.4304403 0.550500 2.42  
## 8 Age 31.2052239 37.067164 81.00  
## 9 Outcome 1.0000000 2.000000 2.00

# 최종 결론

### 모든 변수의 평균이 환자가 높을 것으로 예상한 바와 같았다.

### 정상과 환자의 가장 큰 차이는 포도당 부하 검사 수치와 혈청 인슐린이다.

### 인슐린 데이터에 문제가 있는 것 같다.

### 허벅지 근육이 많으면 포도당을 근육세포에 필요로 하기 때문에 혈당이 높아지지 않으므로 허벅지 운동을 하자.