채널 G
2016125077 최재혁
2016126049 박희재
2016125069 조세희

데이터 사이언스

- 6주차 팀 과제 (Naiive bayes algorithm 을 통한 심장질환 환자 예측)

목 차

- 과제 제목

  - 심장질환 환자 데이터를 이용하여 심장질환 예측 모델만들기(Naivebayes 모델)

- 과제 목표

  - 속성의 상관관계를 따져 속성간의 독립성을 구할 수 있다.
  - Naivebayes model을 적용하여 원하는 주제에 대해 자유롭게 데이터 예측을 할 수 있다.

- 세부 목표

  - 모든 독립변수로 만든 모델과 변수간 상관관계를 통해 추려낸 변수들을 이용한 모델의 정확성을 비교하며 더 나은 모델을 선택한다.

목 차

세부 목표에 따른 과제 진행 순서

- 분석할 데이터 셋 조사
- 데이터 마이닝 및 데이터 추출
- 모델 생성 및 정확도 향상을 위한 방법 모색
- 모델 사용, 목표변수 예측

분석할 데이터 셋 조사

- 데이터 셋 결정 – 각 속성이 독립성을 가질 dataset 조사

  - 코로나로 인해 질병에 대한 관심증가, 사망률 1위 질병에 대한 궁금증으로 이어져 조사하던 중, 심장질환이 전연령대에서 발생하며, 이에 따라 관련 data set 존재파악
  - 새로운 정보 발생시 모델에서 활용하여 최선의 예측이 용이한 특징을 가진 Naivebayes 모델을 만들어 새로운 환자 발생시, 최선의 예측을 위해 해당데이터를 설정하였다
    ※ 참조(https://www.kaggle.com/ronitf/heart-disease-uci)

  - **'심장질환'과 관련 Dataset으로 설정**

```
> library(ggplot2)
> df <- read.csv(file = "C:/Users/wligh/Desktop/Heart.csv",header=TRUE, fileEncoding = "utf-8")
> df
   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1   63   1  3      145  233   1       0     150     0     2.3     0  0    1      1
2   37   1  2      130  250   0       1     187     0     3.5     0  0    2      1
3   41   0  1      130  204   0       0     172     0     1.4     2  0    2      1
4   56   1  1      120  236   0       1     178     0     0.8     2  0    2      1
5   57   0  0      120  354   0       1     163     1     0.6     2  0    2      1
```

<사진 1> dataset을 R에 load한 모습

```
#age - 나이
#sex - (1 = 남성; 0 = 여성)
#cp - 가슴 통증 유형(0, 1, 2, 3, 4)
#trestbps - 안정 혈압(병원 입원시 mm Hg)
#chol - 혈청 콜레스테롤(mg/dl)
#fbs - (공복 혈당 > 120 mg/dl)(1 = true; 0 = false)
#restecg - 안정 심전도 결과(0, 1, 2)
#thalach - 최대 심박동수
#exang - 협심증 유발 운동(1 = yes; 0 = no)
#oldpeak - 비교적 안정되기까지 운동으로 유발되는 ST depression
#slope - 최대 운동 ST segment의 기울기
#ca - 형광 투시된 주요 혈관의 수(0-3)
#thal - (3 = 보통; 6 = 해결된 결함; 7 = 해결가능한 결함)
#target - 심장병 진단(1 = true; 0 = false)
```

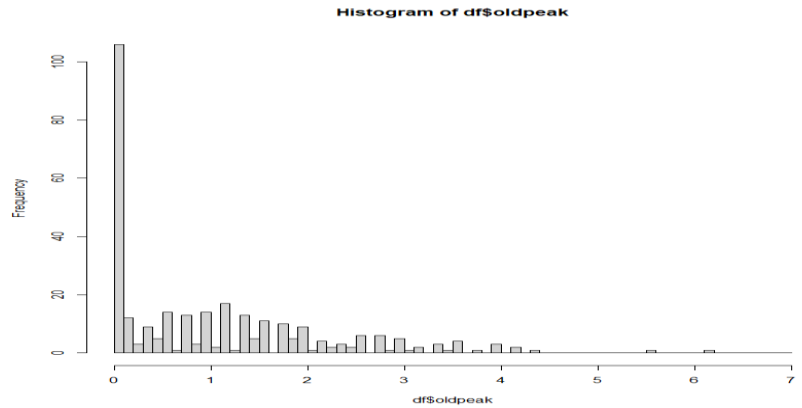
<사진 2> 각 속성 값에 대한 설명

데이터 마이닝 및 추출

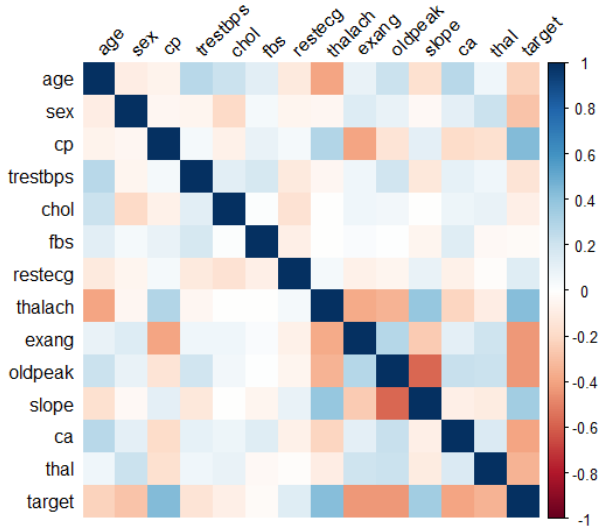
- 데이터 마이닝 – 연관성 파악, data type에 따른 분석

  - 상관계수를 통한 연관성 파악 :명목형 변수 수치화 진행
    각 속성 간 독립적 ∴ 각 속성간 연관성이 적으며, target 변수와의 연관성이 높은 변수 파악
  - 수치형 변수의 경우 : 확률밀도함수가 정규분포표를 따르는 지 분석
    분석 결과 : oldpeak 속성의 경우 따르지 않음.

  - 위 작업 결과를 토대로 분석에 사용할 변수 설정



<사진 1> oldpeak 확률밀도함수graph



<사진 2> 상관계수 시각화 graph

데이터 마이닝 및 추출

- 데이터 속성파악

  변수 특징
  - CP : 가슴 통증 유형
  - SEX : 성별
  - Thalach : 최대 심장박동수
  - Exang : 협심증 유발 운동
  - Slope : 최대 운동 St segment의 기울기
  - Ca : 형광투시된 주요 혈관 (0-3)
  - Thal : (3 = 보통, 6 = 해결된 결함, 7 = 해결가능한 결함)
  - Target : 심장병의 유무(1 = true, 0 = false)
  - Threstbps : 안정 혈압(병원 입원시 mm Hg)
  - Col : 혈청 콜레스테롤(mg/dl)
  - Fbs : 공복혈당(1=true, 0 = false)
  - Restecg : 안정 심전도 결과 (0,1,2)
  - Oldpeak : 비교적 안정되기까지 운동으로 유발되는 ST depression
  - Age : 나이

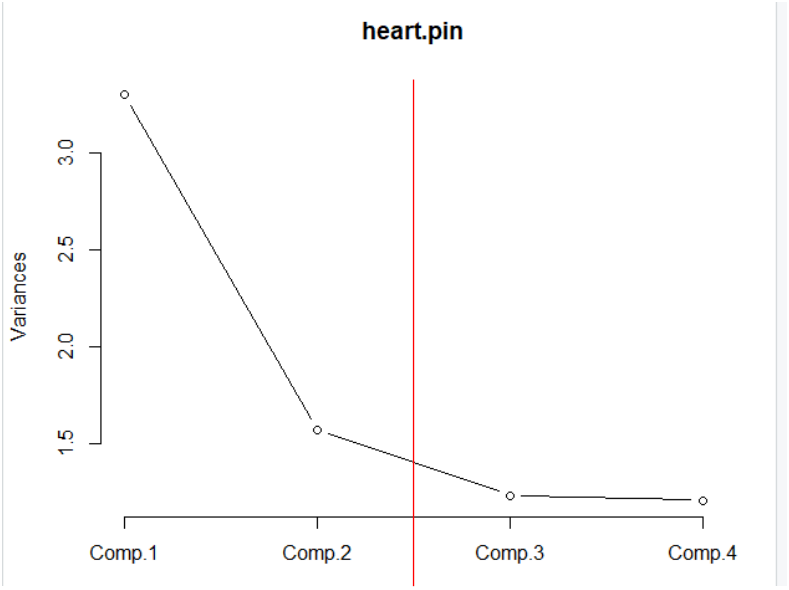
데이터 마이닝 - 주성분분석

- 주성분 분석을 통하여 상관성이 높은 변수들을 파악한다.
  - 서로 상관성이 높은 변수들의 선형 결합으로 만들어 전체 데이터 변수들 중에 상관성이 높은 변수들을 요약, 축소 시켜본다.
  - 주성분분석을 통하여 소수의 주성분들로 차원을 축소시켜 많은 모델을 더욱 단순하고 정확성이 높게 만든다.

```
10  heart.pin <- princomp(df,cor=TRUE)
11  summary(heart.pin)
12  screeplot(heart.pin,npcs = 4,type = "lines")
13  loadings(heart.pin)
14
```

<사진 1> 코드

보통 누적기여율은 85%이상에서 주성분 수로 결정하는데 Screeplot과 비교해보면 일치하지 않는 부분이 있다. 따라서, 주성분 분석은 맞지 않은 데이터 마이닝이라고 판단하였다.



<사진 2> Screeplot

```
> heart.pin <- princomp(df,cor=TRUE)
> summary(heart.pin)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     1.8169962 1.2538599 1.10996801 1.09847117 1.01095452
Proportion of Variance 0.2358197 0.1122975 0.08800207 0.08618849 0.07300207
Cumulative Proportion  0.2358197 0.3481171 0.43611922 0.52230771 0.59530979
                          Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.98497128 0.92910425 0.88096135 0.85392898 0.78912657
Proportion of Variance 0.06929774 0.06165962 0.05543521 0.05208534 0.04448005
Cumulative Proportion  0.66460753 0.72626715 0.78170236 0.83378770 0.87826775
                         Comp.11    Comp.12    Comp.13    Comp.14
Standard deviation     0.73102824 0.65576949 0.60981648 0.60658013
Proportion of Variance 0.03817159 0.03071669 0.02656258 0.02628139
Cumulative Proportion  0.91643934 0.94715603 0.97371861 1.00000000
```

<사진 3> 누적기여율 및 주성분 파악

데이터 마이닝 - 회귀분석

- 선형 회귀 분석을 통한 독립변수 추출
  - 다중 선형 회귀 분석을 통하여 종속변수에 미치는 영향이 가장 큰 독립변수들을 추출하고 이를 통하여 모델을 생성해 본다.

```
15  #회귀분석
16  model0 <- lm(target~age+sex+cp+trestbps+chol+fbs+restecg+thalach+exang+oldpeak+sl
17  summary(model0)
18
```

<사진 1> 코드

```
> model0 <- lm(target~age+sex+cp+trestbps+chol+fbs+restecg+thalach+exang+oldpeak+slope
+ca+thal,data = df)
> summary(model0)

Call:
lm(formula = target ~ age + sex + cp + trestbps + chol + fbs +
    restecg + thalach + exang + oldpeak + slope + ca + thal,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94748 -0.21270  0.06608  0.25022  0.93509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8288987  0.2929344   2.830 0.004987 **
age         -0.0008204  0.0026962  -0.304 0.761129
sex         -0.1959956  0.0471429  -4.157 4.24e-05 ***
cp           0.1127034  0.0223816   5.036 8.40e-07 ***
trestbps    -0.0019910  0.0012573  -1.583 0.114407
chol        -0.0003535  0.0004217  -0.838 0.402545
fbs          0.0173736  0.0596669   0.291 0.771125
restecg      0.0498480  0.0399228   1.249 0.212819
thalach      0.0030193  0.0011304   2.671 0.007988 **
exang       -0.1440459  0.0513689  -2.804 0.005387 **
oldpeak     -0.0587887  0.0229269  -2.564 0.010847 *
slope        0.0789788  0.0423896   1.863 0.063453 .
ca          -0.1006022  0.0218565  -4.603 6.25e-06 ***
thal        -0.1190392  0.0356550  -3.339 0.000952 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3542 on 289 degrees of freedom
Multiple R-squared:  0.5175,    Adjusted R-squared:  0.4958
F-statistic: 23.85 on 13 and 289 DF,  p-value: < 2.2e-16
```

<사진 2> summary

P – value 값은 유의수준 0.05 보다 작으므로 회귀계수의 추정치들이 통계적으로 유의하나, 수정된 결정계수가 0.49로 모델의 설명력은 떨어진다.

데이터 마이닝 – 벌점화 선택

- 벌점화 선택을 통하여 영향력이 떨어지는 변수들을 제거한다.

```
model0 <- step(model0,direction = "backward")
model_lmnew <- lm(target~sex + cp + trestbps + restecg + thalach + exang + oldpe
                  slope + ca + thal,data = df)
summary(model_lmnew)
```

<사진 1> 코드

```
> model0 <- step(model0,direction = "backward")
Start:  AIC=-615.31
target ~ age + sex + cp + trestbps + chol + fbs + restecg + thalach +
    exang + oldpeak + slope + ca + thal

           Df Sum of Sq    RSS     AIC
- fbs       1    0.0106 36.266 -617.22
- age       1    0.0116 36.267 -617.22
- chol      1    0.0882 36.344 -616.58
- restecg   1    0.1956 36.451 -615.68
<none>                  36.255 -615.31
- trestbps  1    0.3146 36.570 -614.70
- slope     1    0.4355 36.691 -613.69
- oldpeak   1    0.8248 37.080 -610.50
- thalach   1    0.8951 37.150 -609.92
- exang     1    0.9865 37.242 -609.18
- thal      1    1.3983 37.654 -605.85
- sex       1    2.1684 38.424 -599.71
- ca        1    2.6578 38.913 -595.88
- cp        1    3.1810 39.436 -591.83

Step:  AIC=-617.22
target ~ age + sex + cp + trestbps + chol + restecg + thalach +
    exang + oldpeak + slope + ca + thal

           Df Sum of Sq    RSS     AIC
- age       1    0.0103 36.276 -619.14
- chol      1    0.0888 36.355 -618.48
- restecg   1    0.1916 36.458 -617.63
<none>                  36.266 -617.22
- trestbps  1    0.3045 36.571 -616.69
- slope     1    0.4276 36.694 -615.67
- oldpeak   1    0.8464 37.112 -612.23
- thalach   1    0.9038 37.170 -611.77
- exang     1    0.9781 37.244 -611.16
- thal      1    1.4160 37.682 -607.62
- sex       1    2.1582 38.424 -601.71
```

<사진 2> 벌점화 선택

P – value 값은 여전히 유의하나, 수정된 결정계수는 올랐음에도 불구하고 아직 많이 부족하다. 하지만 모델을 만들어보고 정확도를 확인해보겠다.

```
> summary(model_lmnew)

Call:
lm(formula = target ~ sex + cp + trestbps + restecg + thalach +
    exang + oldpeak + slope + ca + thal, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95600 -0.20876  0.05823  0.25634  0.91861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.705308   0.235654   2.993 0.002999 **
sex         -0.183737   0.045307  -4.055 6.43e-05 ***
cp           0.113976   0.022106   5.156 4.66e-07 ***
trestbps    -0.002112   0.001201  -1.758 0.079844 .
restecg      0.055506   0.039193   1.416 0.157775
thalach      0.003118   0.001046   2.981 0.003119 **
exang       -0.144356   0.050915  -2.835 0.004899 **
oldpeak     -0.060192   0.022748  -2.646 0.008585 **
slope        0.076055   0.042067   1.808 0.071646 .
ca          -0.101947   0.021128  -4.825 2.26e-06 ***
thal        -0.123490   0.035234  -3.505 0.000529 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.353 on 292 degrees of freedom
Multiple R-squared:  0.5159,    Adjusted R-squared:  0.4993
F-statistic: 31.12 on 10 and 292 DF,  p-value: < 2.2e-16
```

<사진 3> 벌점화 선택후 모델 summary

데이터 분할 및 기본 모델

- 데이터 셋을 균일하게 분할한다

```
33 #데이터 분할
34 set.seed(1234)
35 intrain<-createDataPartition(y=df$target, p=0.7, list=FALSE)
36 train<-df[intrain, ]
37 test<-df[-intrain, ]
38 print(table(train$target))
39 print(table(test$target))
40
```

<사진 1> 데이터 분할

- 분할된 데이터를 가지고 전체변수를 통한 모델 생성 및 정확도 확인

```
42 #나이브 베이즈 모델 생성
43 model <- naiveBayes(target~.,data = train)
44 model
45 summary(model)
46
47 #예측
48 pred <- predict(model, test, type='class')
49 pred
50 confusionMatrix(pred, test$target)
```

<사진 2> 베이지안 모델 생성 및 예측 코드

```
> confusionMatrix(pred, test$target)
Confusion Matrix and Statistics

          Reference
Prediction zero one
      zero   32   5
      one     9  44

               Accuracy : 0.8444
                 95% CI : (0.7528, 0.9123)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 1.629e-09

                  Kappa : 0.6839

 Mcnemar's Test P-Value : 0.4227

            Sensitivity : 0.7805
            Specificity : 0.8980
         Pos Pred Value : 0.8649
         Neg Pred Value : 0.8302
             Prevalence : 0.4556
         Detection Rate : 0.3556
   Detection Prevalence : 0.4111
      Balanced Accuracy : 0.8392

       'Positive' Class : zero
```

현재 전체 독립변수를 가지고 나이브 베이지안 모델을 생성했을 시,
84%의 정확도를 띈다.

회귀분석을 통한 모델 생성

- 회귀분석을 통해 추려낸 독립 변수로 모델을 생성
  - 다중 선형 회귀 분석을 통하여 유의한 독립변수를 추려보았고, 정확도를 확인하였는데, 85%로 1%상승한 것을 확인할 수 있다.
  - 이로써 회귀분석이 성공적이었다는 것을 알 수 있다.

```
42 #나이브 베이즈 모델 생성
43 model <- naiveBayes(target~sex + cp + trestbps + restecg + thalach +
44                       exang + oldpeak + slope + ca + thal,data = train)
45 model
46 summary(model)
47
48 #예측
49 pred <- predict(model, test, type='class')
50 pred
51 confusionMatrix(pred, test$target)
```

<사진 1> 코드 (동일한 테스트 데이터를 통해서 모델을 예측)

```
> confusionMatrix(pred, test$target)
Confusion Matrix and Statistics

          Reference
Prediction zero one
      zero   32   4
      one     9  45

               Accuracy : 0.8556
                 95% CI : (0.7657, 0.9208)
    No Information Rate : 0.5444
    P-Value [Acc > NIR] : 3.463e-10

                  Kappa : 0.7059

 Mcnemar's Test P-Value : 0.2673

            Sensitivity : 0.7805
            Specificity : 0.9184
         Pos Pred Value : 0.8889
         Neg Pred Value : 0.8333
             Prevalence : 0.4556
         Detection Rate : 0.3556
   Detection Prevalence : 0.4000
      Balanced Accuracy : 0.8494

       'Positive' Class : zero
```

<사진 2> 다중 선형 회귀분석을 통해 구한 혼동 행렬

데이터 마이닝 - 상관계수

- Cor()함수를 통하여 상관계수를 확인
  - 음, 양의 선형관계를 갖는 변수들을 확인하여 이를 제거하고, 나이브 베이지안 모델을 생성한 후, 모델을 생성해 본다.

```
> cor(df) #상관계수
                age         sex          cp    trestbps        chol         fbs     restecg     thalach       exang     oldpeak
age      1.00000000 -0.09844660 -0.06865302  0.27935091  0.213677957  0.12130765 -0.11621090 -0.39852194  0.09680083  0.21001257
sex     -0.09844660  1.00000000 -0.04935288 -0.05676882 -0.19791217  0.04503179 -0.05819627 -0.04401991  0.14166381  0.09609288
cp      -0.06865302 -0.04935288  1.00000000  0.04760776 -0.07690439  0.09444403  0.04442059  0.29576212 -0.39428027 -0.14923016
trestbps 0.27935091 -0.05676882  0.04760776  1.00000000  0.12317421  0.17753054 -0.11410279 -0.04669773  0.06761612  0.19321647
chol     0.21367796 -0.19791217 -0.07690439  0.12317421  1.000000000 0.01329360 -0.15104008 -0.00993983  0.06702278  0.05395192
fbs      0.12130765  0.04503179  0.09444403  0.17753054  0.013293602 1.00000000 -0.08418905 -0.00856711  0.02566515  0.00574722
restecg -0.11621090 -0.05819627  0.04442059 -0.11410279 -0.151040078 -0.08418905  1.00000000  0.04412344 -0.07073286 -0.05877023
thalach -0.39852194 -0.04401991  0.29576212 -0.04669773 -0.009939839 -0.00856711  0.04412344  1.00000000 -0.37881209 -0.34418695
exang    0.09680083  0.14166381 -0.39428027  0.06761612  0.067022783  0.02566515 -0.07073286 -0.37881209  1.00000000  0.28822281
oldpeak  0.21001257  0.09609288 -0.14923016  0.19321647  0.053951920  0.00574722 -0.05877023 -0.34418695  0.28822281  1.00000000
slope   -0.16881424 -0.03071057  0.11971659 -0.12147458 -0.004037770 -0.05989418  0.09304482  0.38678441 -0.25774837 -0.57753682
ca       0.27632624  0.11826141 -0.18105303  0.10138899  0.070510925  0.13797933 -0.07204243 -0.21317693  0.11573938  0.22268232
thal     0.06800138  0.21004110 -0.16173557  0.06220989  0.098802993 -0.03201934 -0.01198140 -0.09643913  0.20675379  0.21024413
target  -0.22543872 -0.28093658  0.43379826 -0.14493113 -0.085239105 -0.02804576  0.13722950  0.42174093 -0.43675708 -0.43069600
              slope          ca        thal      target
age      -0.16881424  0.27632624  0.06800138 -0.22543872
sex      -0.03071057  0.11826141  0.21004110 -0.28093658
cp        0.11971659 -0.18105303 -0.16173557  0.43379826
trestbps -0.12147458  0.10138899  0.06220989 -0.14493113
chol     -0.00403777  0.07051093  0.09880299 -0.08523911
fbs      -0.05989418  0.13797933 -0.03201934 -0.02804576
restecg   0.09304482 -0.07204243 -0.01198140  0.13722950
thalach   0.38678441 -0.21317693 -0.09643913  0.42174093
exang    -0.25774837  0.11573938  0.20675379 -0.43675708
oldpeak  -0.57753682  0.22268232  0.21024413 -0.43069600
slope     1.00000000 -0.08015521 -0.10476379  0.34587708
ca       -0.08015521  1.00000000  0.15183213 -0.39172399
thal     -0.10476379  0.15183213  1.00000000 -0.34402927
target    0.34587708 -0.39172399 -0.34402927  1.00000000
```

전체적으로 변수들을 비교해 보았을 때 강한 음, 양의 상관관계를 갖는 변수들은 확실히 구분되지는 않으나, 비교적 높은 상관관계를 갖는 변수 thalach, oldpeak, slope 중, slope를 제거하고 나머지 변수들은 제거하여 모델을 생성해본다

모델에 새로운 데이터 적용

- 새로운 데이터를 생성하여 모델에 넣어보고, 결과를 확인해본다.
  - 정확도가 가장 높았던, 마지막 모델을 가지고 새로운 데이터 2명을 만들고 예측해 보았다.

```
52 #판별별
53 target <- data.frame(age=24,sex=1, cp=0, trestbps=130,chol=120,fbs=0,restecg=0,thalach=170,exang=1,oldpeak=0.2,slope=1,ca=0,thal=2)
54 target2 <- data.frame(age=43,sex=1, cp=0, trestbps=120,chol=177,fbs=0,restecg=0,thalach=120,exang=1,oldpeak=2.5,slope=1,ca=0,thal=3)
55
56 predict(model, newdata=target)
57 predict(model, newdata=target2)
```

<사진 1> 사람1 – target, 사람2 – target2

```
> #판별별
> target <- data.frame(age=24,sex=1, cp=0, trestbps=130,chol=120,fbs=0,restecg=0,thalach=170,exang=1,oldpeak=0.2,slope=1,ca=0,thal=2)
> target2 <- data.frame(age=43,sex=1, cp=0, trestbps=120,chol=177,fbs=0,restecg=0,thalach=120,exang=1,oldpeak=2.5,slope=1,ca=0,thal=3)
> predict(model, newdata=target)
[1] one
Levels: zero one
> predict(model, newdata=target2)
[1] zero
Levels: zero one
```

<사진 2> 첫번째사람 – 심장병(true), 두번째사람 – 심장병(false)

결론 및 고찰

- 결론 및 고찰

  - 각 개체의 독립성 파악 과정에서 다양한 방법으로 추출해보고 각 상황에 맞게 이상적인 추출을 해야 naivebayes 모델 생성시 정확도를 높이는 가장 큰 요인이 될 것 같다.

  - 주성분분석, 회귀분석, 상관계수 파악 등 다양한 데이터 마이닝을 통하여 모델 향상을 위한 데이터 분석을 해보았다는 것에 큰 흥미가 생겼다.

  - 나이브 베이지안 모델과 지난 주, 의사결정나무 모델 두가지를 배워보면서 분류 기법에 대해서 다양하게 접근할 수 있게 되어 보람을 느꼈다.