# 데이터마이닝
# 프로젝트 보고서


| 제출일 | 2020.06,15 | 전공 | 컴퓨터과학과 |
|---|---|---|---|
| 과목 | 데이터마이닝 | 학번 | 201411701 |
| 담당교수 | 정경용 | 이름 | 김태균 |

# 1. 데이터 수집

## 1-1. 분석 대상 데이터

Heart Disease UCI - 환자의 정보를 이용한 심장 질환 판단 여부
출처:https://www.kaggle.com/ronitf/heart-disease-uci

## 1-2. 데이터 가져오기

```
heart=read.csv("C:/Users/ktg73/Desktop/4-1/데이터마이닝/프로젝트
/HeartDiseaseUCI.csv")
```

## 1-3. 데이터 구조 파악 및 분석

```
> str(heart)
'data.frame':    303 obs. of  14 variables:
 $ age      : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex      : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cp       : int  3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol     : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs      : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang    : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope    : int  0 0 2 2 2 1 1 2 2 2 ...
 $ ca       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thal     : int  1 2 2 2 2 1 2 3 3 2 ...
 $ target   : int  1 1 1 1 1 1 1 1 1 1 ...
```

age: 나이
sex: 성별(1=male; 0=female)
cp: 흉통유형(0=전형적 협심증; 1:비정형 협심증; 2:비 협심증; 3:무증상
trestbps: 휴식시 혈압(mmHg)
chol: 혈청 콜레스테롤(mg/dl)
fbs: 공복혈당>120mg/dl (1=True; 0=False)
restecg: 휴식시 심전도 결과(0=정상; 1=ST-T파 이상(T파 반전 또는 ST가 0.05mV 상승/감소; 2=Estes기준에 의거 명확하게 좌심실 비대가 보임)
thalach:최대 심박수
exang: 활동 유발 협심증(1=True; 0=False)
oldpeak: 휴식과 관련된 활동으로 인한 ST depression
slope: peak운동 ST세그먼트의 기울기(0=상승; 1=평면; 2=감소)
ca: 투시검사에 의해 착색되는 주요 혈관 수(주요혈관은 0~3이 존재, 즉 최대 4)
thal: 탈륨 심장 스캔(1= 정상(냉점 없음); 2= 고정 결함(휴식 활동 중 냉점); 3= 가역적 결함(활동 중 냉점이 나타날 때);
target: 심장병 진단(혈관 질환 상태)(1=심장병 존재; 0=심장병 없음)

## 1-4. 데이터 샘플 해석

```
> head(heart,2)
  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
```

첫 번째 샘플의 경우 흉통 무증상, 휴식시 혈압이 145mmHg, 혈청 콜레스테롤이 233mg/dl. 공복혈당>120mg/dl, 휴식시 심전도 결과가 정상, 최대 심박수 150, 활동 유발 협심증 없음,

ST우울증 없음, ST세그먼트 기울기 상승, 투시검사에 의해 착색된 주요 혈관 수=0, 탈륨 심장 스캔 결과 정상인 63세의 남성은 심장병이 존재한다.

# 2. 데이터 정제

## 2-1. 결측값 처리

```
> table(is.na(heart))

FALSE
 4242
```

결측값이 없음

## 2-2. 이상값 처리

이상값이 존재할 수 있는 속성
sex(0~1), cp(1~4), fbs(0~1), restecg(0~2), exang(0~1), slope(1~3),
ca(0~3), thal(1~3), target(0~1)

```
> heart$sex=ifelse((heart$sex<0|heart$sex>1),NA,heart$sex)
> heart$cp=ifelse((heart$cp<0|heart$cp>3),NA,heart$cp)
> heart$fbs=ifelse((heart$fbs<0|heart$fbs>1),NA,heart$fbs)
> heart$restecg=ifelse((heart$restecg<0|heart$restecg>2),NA,heart$restecg)
> heart$exang=ifelse((heart$exang<0|heart$exang>1),NA,heart$exang)
> heart$slope=ifelse((heart$slope<0|heart$slope>2),NA,heart$slope)
> heart$ca=ifelse((heart$ca<0|heart$ca>4),NA,heart$ca)
> heart$thal=ifelse((heart$thal<1|heart$thal>3),NA,heart$thal)
> heart$target=ifelse((heart$target<0|heart$target>1),NA,heart$target)
> table(is.na(heart))

FALSE  TRUE
 4240     2
```

2개의 이상값 검증

```
>
heart=heart[!is.na(heart$sex)&!is.na(heart$cp)&!is.na(heart$fbs)&!is.na(heart$restecg)&!is.na(heart$exang)&!is.na(heart$slope)&!is.na(heart$ca)&!is.na(heart$thal)&!is.na(heart$target),]
```

이상값 포함된 행 제거

```
> str(heart)
'data.frame':   301 obs. of  14 variables:
```

기존 303개의 데이터 중 301개 데이터 남음

# 3. 데이터 가공 및 분석

## 3-1. 각 속성별 개별 영향도 판단

각 속성별 개별 값들이 target에 영향을 주는 정도를 판단, 개별이기에 복합적으로 적용하는 영향에 대해서는 무시한다.

```
> install.packages("dplyr")
> library(dplyr)
```

### 3-1-1. 성별과 심장병

```
> s=select(heart,sex,target)
> table(s[s$target==1,])
    target
sex   1
  0  71
  1  93
> table(s[s$target==0,])
    target
sex    0
  0   24
  1  113
```

성별은 심장병과 큰 연관이 없다고 판단된다.

### 3-1-2. 나이과 심장병

```
> s=select(heart,age,target)
> table(floor(s[s$target==1,]/10)) #몇 십대인지로 구분
    target
age   0
  2   1
  3  11
  4  50
  5  64
  6  32
  7   6
> table(floor(s[,1]/10))
  2   3   4   5   6   7
  1  15  72 123  80  10
```

40, 50, 60대에 심장병이 많은 것 같지만 데이터자체에 40, 50, 60대가 많았기에 판단근거로는 부적합하다.

### 3-1-3. cp와 심장병

```
> s=select(heart,cp,target)
> table(s[s$target==1,])
    target
cp    1
  0  39
  1  41
  2  68
  3  16
> table(s[s$target==0,])
    target
cp     0
  0  103
  1    9
  2   18
```

```
  3   7
```
cp의 경우 1,2,3일 때(흉통이 존재할 경우), 심장병이 있을 확률이 크다.

**3-1-4. 혈압과 심장병**
```
> s=select(heart,trestbps,target)
> mean(s[s$target==1,]$trestbps)
[1] 129.311
> mean(s[s$target==0,]$trestbps)
[1] 134.4453
```
혈압은 심장병과 큰 연관이 없다고 판단된다.

**3-1-5. 혈청 콜레스테롤과 심장병**
```
> s=select(heart,chol,target)
> mean(s[s$target==1,]$chol)
[1] 242.3902
> mean(s[s$target==0,]$chol)
[1] 251.4307
```
혈청 콜레스테롤과 심장병은 큰 연관이 없다고 판단된다.

**3-1-6. 공복 혈당과 심장병**
```
> s=select(heart,fbs,target)
> table(s[s$target==1,])
    target
fbs   1
  0 141
  1  23
> table(s[s$target==0,])
    target
fbs   0
  0 116
  1  21
```
공복 혈당과 심장병은 큰 연관이 없다고 판단된다.

**3-1-7. 심전도 결과와 심장병**
```
> s=select(heart,restecg,target)
> table(s[s$target==1,])
         target
restecg   1
       0 67
       1 96
       2  1
> table(s[s$target==0,])
         target
restecg   0
       0 79
       1 55
       2  3
```
심전도 결과는 ST-T파에 이상이 있을 경우 심장병일 확률이 존재할 수 있다고 판단된다.

**3-1-8. 심박수와 심장병**
```
> s=select(heart,thalach,target)
> mean(s[s$target==1,]$thalach)
[1] 158.7317
> mean(s[s$target==0,]$thalach)
[1] 138.9781
```

최대 심박수가 높으면 심장병일 확률이 존재할 수 있다고 판단된다.

### 3-1-9. 협심증과 심장병

```
> s=select(heart,exang,target)
      target
exang    1
     0 141
     1  23
> table(s[s$target==0,])
      target
exang  0
     0 62
     1 75
```

활동 유발 협심증이 있으면 심장병일 확률이 존재할 수 있다고 판단된다.

### 3-1-10. ST depression과 심장병

```
> s=select(heart,oldpeak,target)
> table(s[s$target==1,]$oldpeak!=0)
FALSE  TRUE
   73    91
> table(s[s$target==0,]$oldpeak!=0)
FALSE  TRUE
   25   112
```

활동으로 인한 ST depression이 0일 때, 심장병일 확률이 존재할 수 있다고 판단된다.

### 3-1-11. ST세그먼트의 기울기와 심장병

```
> s=select(heart,slope,target)
> table(s[s$target==1,]$slope)
  0   1   2
  9  49 106
> table(s[s$target==0,]$slope)
 0  1  2
12 90 35
```

peak운동 ST세그먼트의 기울기가 하락세일 때, 심장병일 확률이 존재할 수 있다고 판단된다.

### 3-1-12. 투시검사와 심장병

```
> s=select(heart,ca,target)
> table(s[s$target==1,]$ca)
  0   1   2   3   4
129  21   7   3   4
> table(s[s$target==0,]$ca)
 0  1  2  3  4
44 44 31 17  1
```

투시검사에 착색되는 주요 혈관수가 0이면 심장병일 확률이 높다.

### 3-1-13. 탈륨심장스캔과 심장병

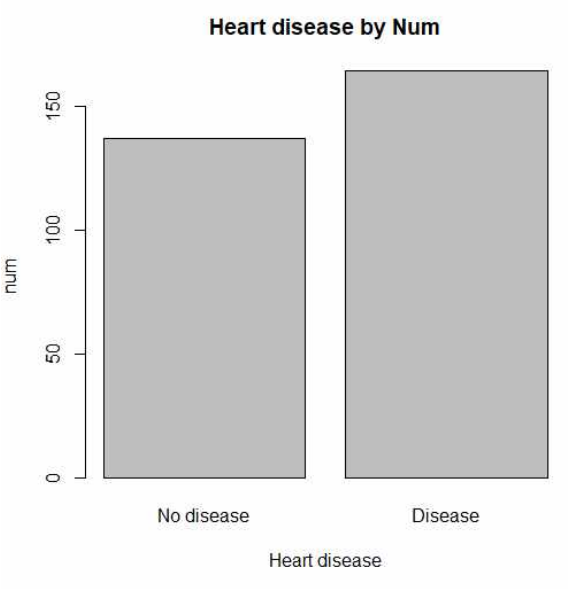
```
> s=select(heart,thal,target)
> table(s[s$target==1,]$thal)
  1   2   3
  6 130  28
> table(s[s$target==0,]$thal)
  1  2  3
 12 36 89
```

탈륨심장스캔 결과 고정결함일 경우 심장병일 확률이 높다.

# 4. 데이터 시각화

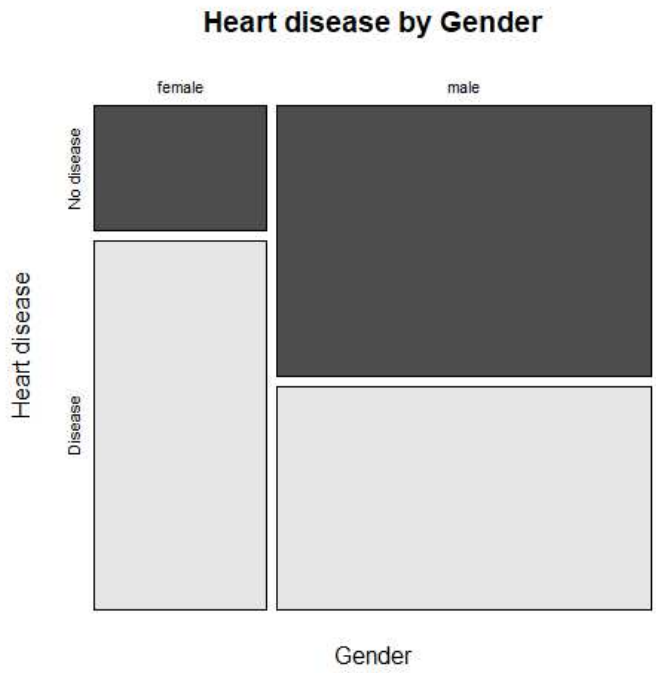
## 4-1. 심장질환유무 환자 파악

```
> data=heart
> data$target=as.factor(data$target)
> levels(data$target) = c("No disease","Disease")
> barplot(table(data$target),main="Heart disease by Num",xlab="Heart disease",
ylab="num")
```



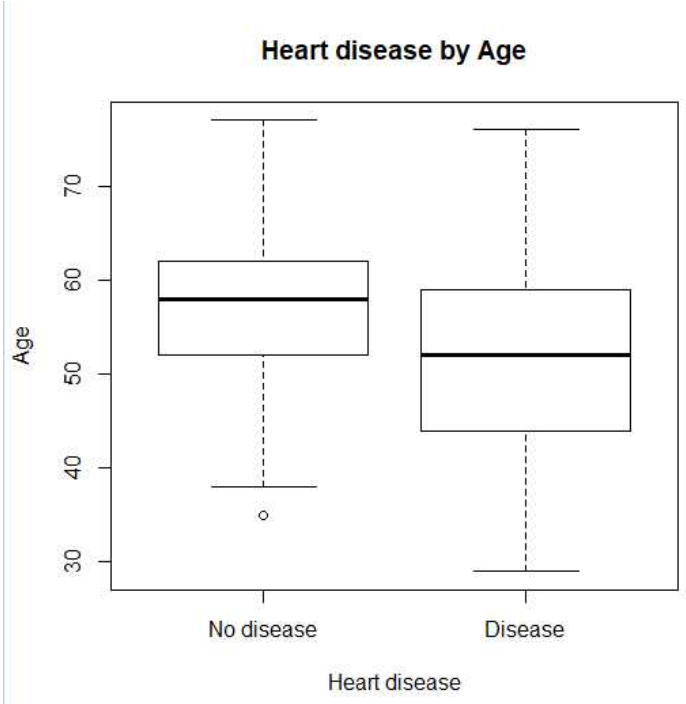

## 4-2. 성별과 심장질환 시각화

```
> data$sex=as.factor(data$sex)
> levels(data$sex) = c("female","male")
> mosaicplot(data$sex ~ data$target, main="Heart disease by Gender",xlab="Gender",
ylab="Heart disease",color=TRUE,shade=FALSE)
```



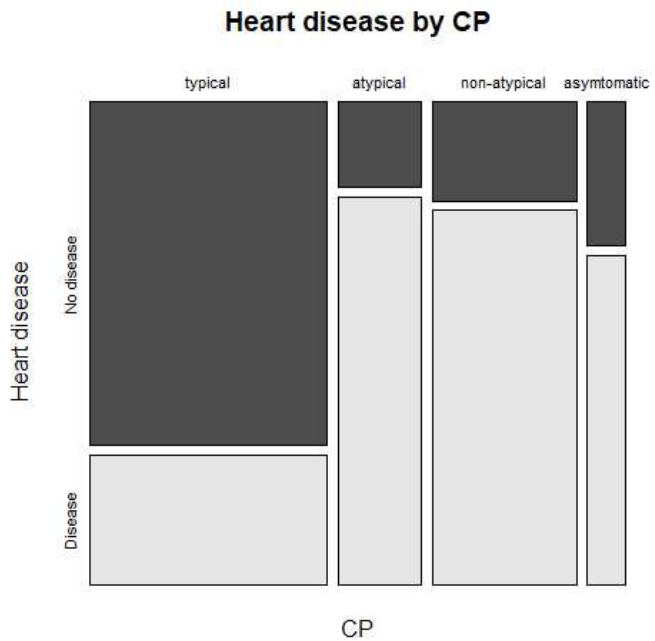
## 4-3. 나이와 심장질환 시각화

> boxplot(data$age~data$target,main="Heart disease by Age",ylab="Age",xlab="Heart disease")



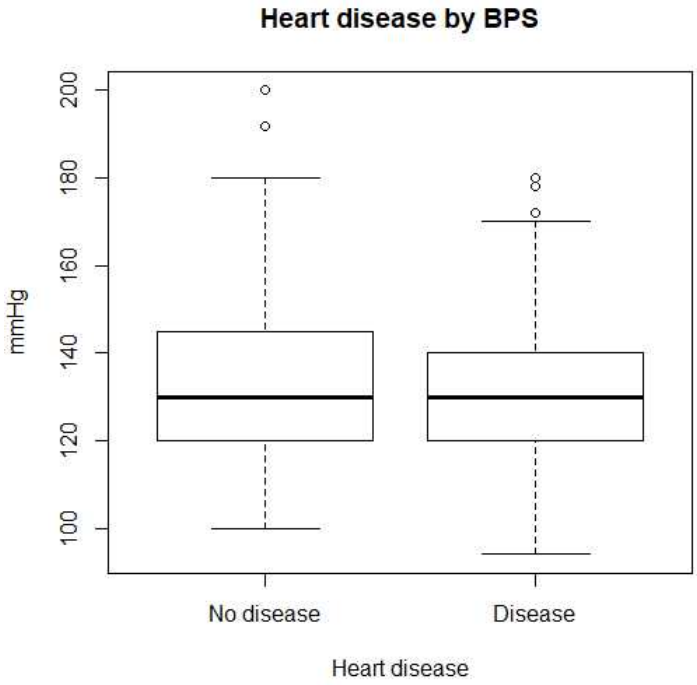

## 4-4. CP와 심장질환 시각화

> data$cp=as.factor(data$cp)
> levels(data$cp) = c("typical","atypical","non-atypical","asymtomatic")
> mosaicplot(data$cp ~ data$target, main="Heart disease by CP",xlab="CP", ylab="Heart disease",color=TRUE,shade=FALSE)



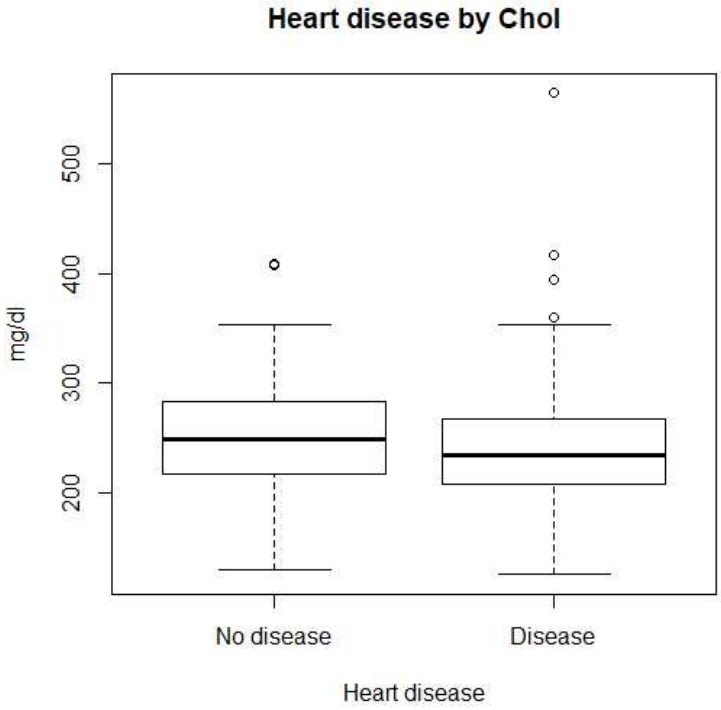
## 4-5. 혈압과 심장질환 시각화

```
> boxplot(data$trestbps~data$target,main="Heart disease by BPS",ylab="BPS",
xlab="Heart disease")
```



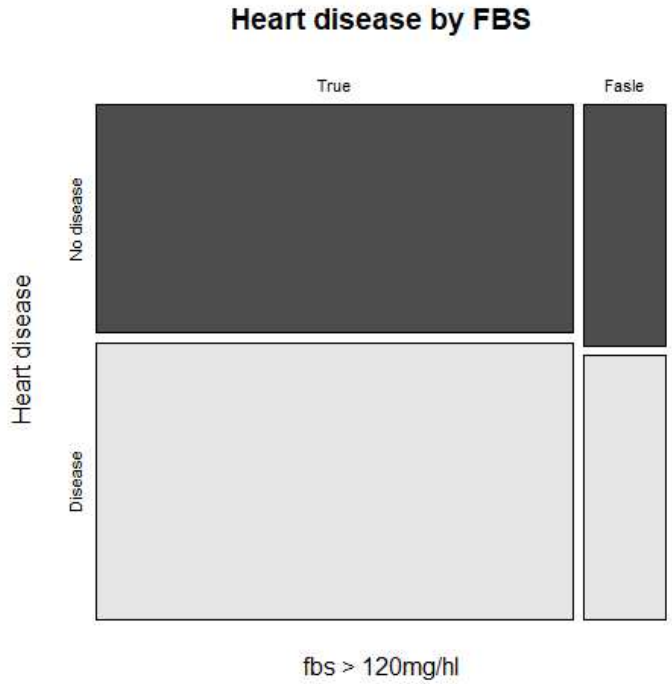

## 4-6. 혈청 콜레스테롤과 심장질환 시각화

```
boxplot(data$chol~data$target,main="Heart disease by Chol",ylab="mg/dl",xlab="Heart
disease")
```



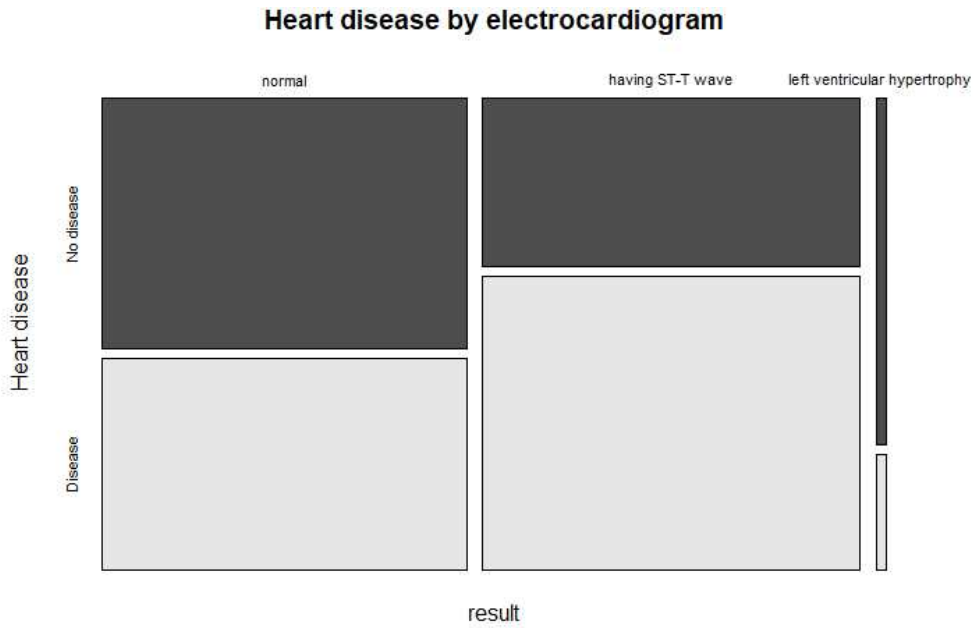
## 4-7. 공복 혈당과 심장질환 시각화

```
> data$fbs=as.factor(data$fbs)
> levels(data$fbs) = c("True","Fasle")
> mosaicplot(data$fbs ~ data$target, main="Heart disease by FBS",xlab="fbs > 120mg/hl", ylab="Heart disease",color=TRUE,shade=FALSE)
```



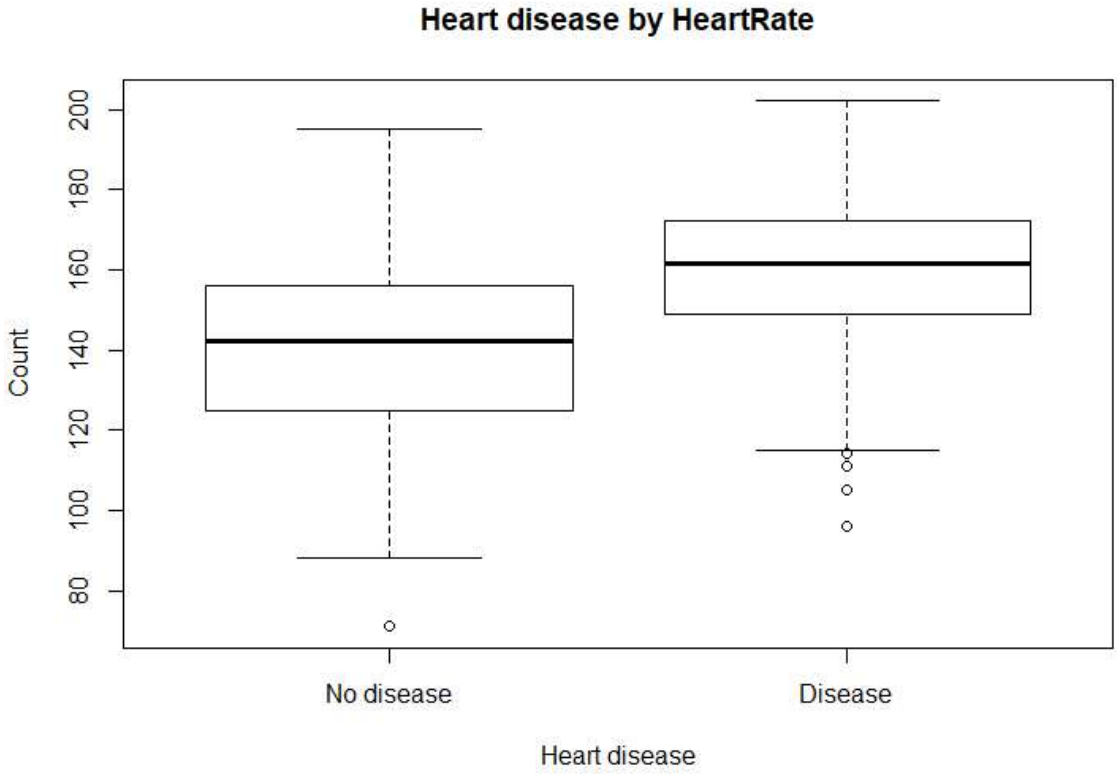

## 4-8. 심전도 결과와 심장질환 시각화

```
> data$restecg=as.factor(data$restecg)
> levels(data$restecg) = c("normal","having ST-T wave","left ventricular hypertrophy")
> mosaicplot(data$restecg ~ data$target, main="Heart disease by electrocardiogram",xlab="result", ylab="Heart disease",color=TRUE,shade=FALSE)
```



## 4-9. 심박수와 심장질환 시각화

```
> boxplot(data$thalach~data$target,main="Heart disease by HeartRate",
ylab="Count",xlab="Heart disease")
```

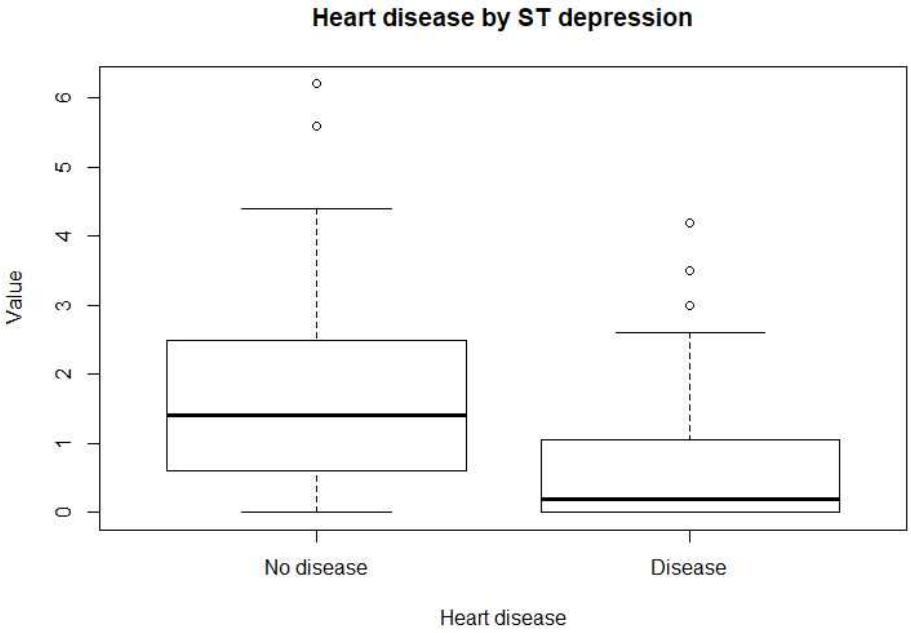



## 4-10. 협심증과 심장질환 시각화

```
> data$exang=as.factor(data$exang)
> levels(data$exang) = c("True","Fasle")
> mosaicplot(data$exang ~ data$target, main="Heart disease by angina",xlab="result",
ylab="Heart disease",color=TRUE,shade=FALSE)
```

## 4-11. ST depression과 심장질환 시각화

```
> boxplot(data$oldpeak~data$target,main="Heart disease by ST depression",
ylab="Value",xlab="Heart disease")
```

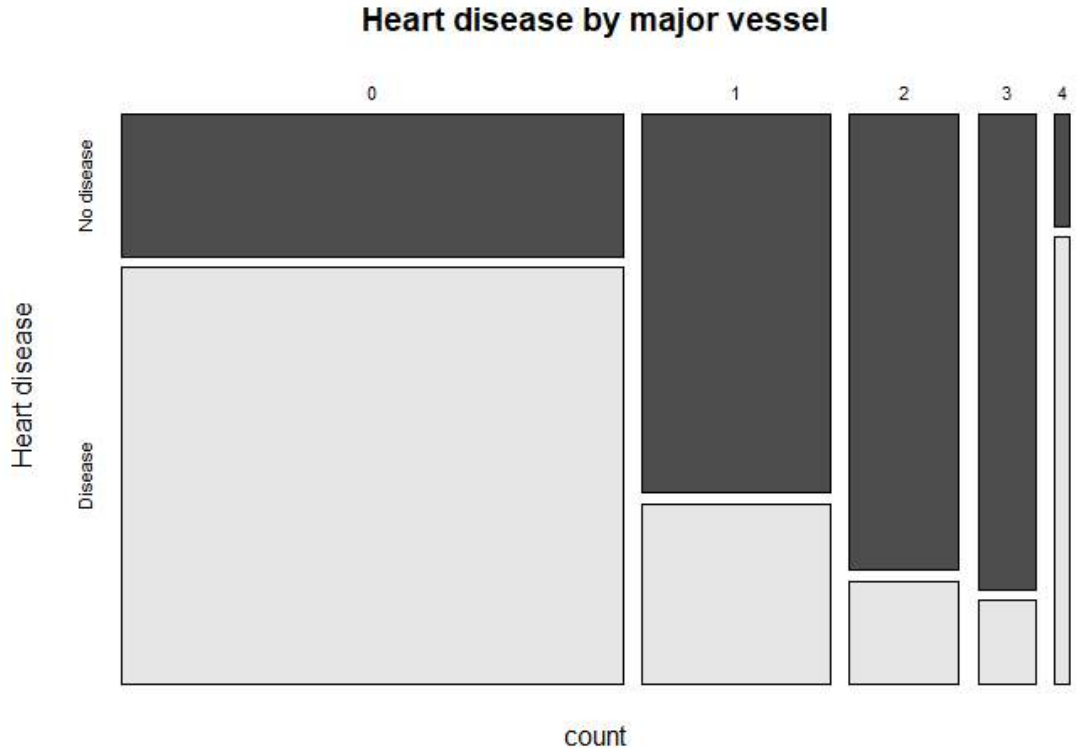



## 4-12. ST세그먼트와 심장질환 시각화

```
> data$slope=as.factor(data$slope)
> levels(data$slope) = c("upsloping","flat","dawnsloping")
> mosaicplot(data$slope ~ data$target, main="Heart disease by ST segment“,
xlab="slope", ylab="Heart disease",color=TRUE,shade=FALSE)
```

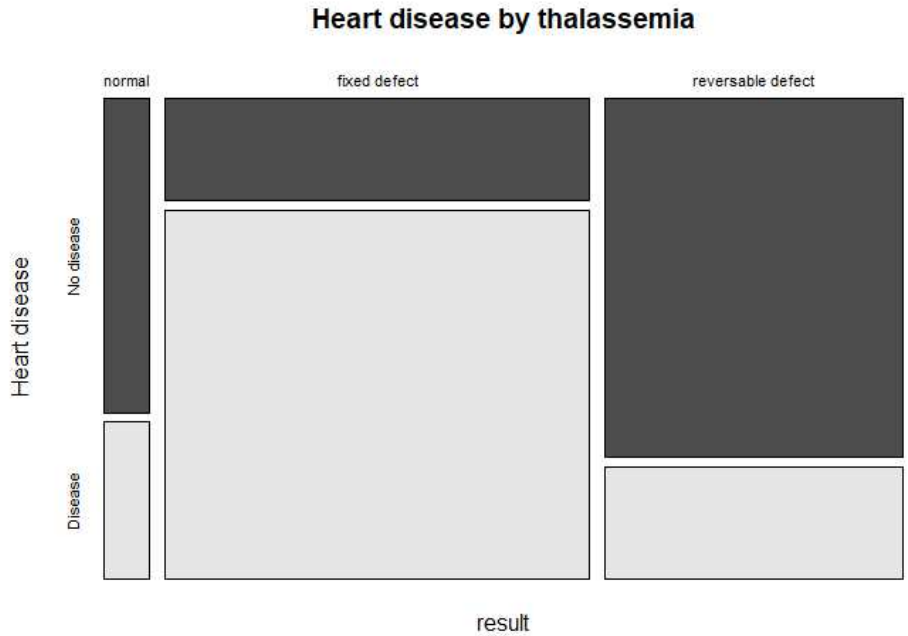
## 4-13. 투시검사와 심장질환 시각화

> mosaicplot(data$ca ~ data$target, main="Heart disease by major vessel",xlab="count", ylab="Heart disease",color=TRUE,shade=FALSE)




## 4-14. 탈륨심장스캔과 심장질환 시각화

> data$thal=as.factor(data$thal)
> levels(data$thal) = c("normal","fixed defect","reversable defect")
> mosaicplot(data$thal ~ data$target, main="Heart disease by thalassemia",xlab="result", ylab="Heart disease",color=TRUE,shade=FALSE)



# 5. 데이터 모델링

## 5-1. 로지스틱 회귀 모델

```
> m=glm(target~.,data=heart,family=binomial)
> coef(m)
 (Intercept)          age          sex           cp      trestbps
 3.450472415 -0.004908470 -1.758180738  0.859850938 -0.019476620
        chol          fbs      restecg      thalach        exang
-0.004630231  0.034887645  0.466282248  0.023210935 -0.979980686
     oldpeak        slope           ca         thal
-0.540273946  0.579288142 -0.773349275 -0.900431861
> deviance(m)
[1] 211.436
```

test case는 기존 heart데이터에서 1과 0이 비슷하게 나뉘는 구간인 165번째를 기준으로 50개의 데이터를 추출해서 사용한다.

```
> new_patients=heart[140:190,1:13]
> predict(m,newdata=new_patients,type='response')
        140         141         142         143         144         145
0.031904907 0.980050393 0.653851661 0.992748753 0.807789186 0.951079639
        146         147         148         149         150         151
0.670990080 0.964536678 0.993370580 0.972321670 0.958753609 0.357783816
        152         153         154         155         156         157
0.819404816 0.678554648 0.904526387 0.982793100 0.843280182 0.969428495
        158         159         160         161         162         163
0.951565284 0.056859369 0.723659974 0.804849584 0.948562768 0.965614824
        164         165         166         167         168         169
0.616211109 0.616211109 0.002725660 0.006100401 0.083151307 0.071444169
        170         171         172         173         174         175
0.018742080 0.587107738 0.559204589 0.533686900 0.216940387 0.006898659
        176         177         178         179         180         181
0.049443665 0.071971184 0.909810192 0.034129133 0.070048706 0.022387726
        182         183         184         185         186         187
0.040381815 0.914626438 0.369288635 0.040875454 0.521937578 0.072466692
        188         189         190
0.008548142 0.596557770 0.663029642
```

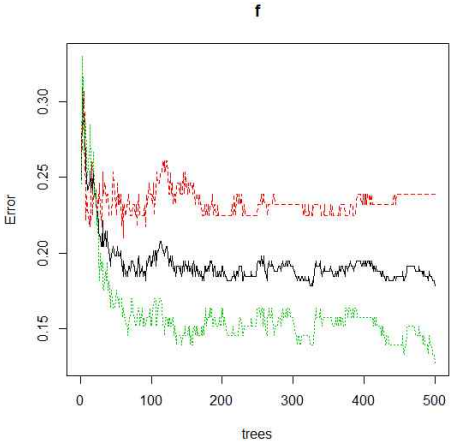
```
> mean(predict(m,newdata=new_patients,type='response'))
[1] 0.5232982
```

50개의 데이터에 대해 52.32%의 정확률을 가진다.

```
> m
Call:  glm(formula = target ~ ., family = binomial, data = heart)

Coefficients:
(Intercept)          age          sex           cp      trestbps
   3.450472    -0.004908    -1.758181     0.859851    -0.019477
       chol          fbs      restecg      thalach        exang
  -0.004630     0.034888     0.466282     0.023211    -0.979981
    oldpeak        slope           ca         thal
  -0.540274     0.579288    -0.773349    -0.900432

Degrees of Freedom: 302 Total (i.e. Null);  289 Residual
Null Deviance:      417.6
Residual Deviance: 211.4        AIC: 239.4
```

## 5-2. 랜덤 포리스트 모델

```
> install.packages("randomForest")
> library(randomForest)
> heart$target=as.factor(heart$target)
> f=randomForest(target~.,data=heart)
> f
Call:
 randomForest(formula = target ~ ., data = heart)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 17.82%
Confusion matrix:
    0   1 class.error
0 105  33   0.2391304
1  21 144   0.1272727
> summary(f)
                Length Class  Mode
call              3    -none- call
type              1    -none- character
predicted       303    factor numeric
err.rate       1500    -none- numeric
confusion         6    -none- numeric
votes           606    matrix numeric
oob.times       303    -none- numeric
classes           2    -none- character
importance       13    -none- numeric
importanceSD      0    -none- NULL
localImportance   0    -none- NULL
proximity         0    -none- NULL
ntree             1    -none- numeric
mtry              1    -none- numeric
forest           14    -none- list
y               303    factor numeric
test              0    -none- NULL
inbag             0    -none- NULL
terms             3    terms  call
> plot(f)
```




randomForest의 평균오류율은 17.82로 정확률은 82.28%이다.

## 5-3. SVM모델

### 5-3-1. radial basis function커널

```
> library(e1071)
> s=svm(target~.,data=heart)
> print(s)

Call:
svm(formula = target ~ ., data = heart)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  174

> table(predict(s,heart),heart$target)

     0   1
  0 124  10
  1  14 155
```

오류율: 24/303

### 5-3-2. ploynomial커널

```
> s=svm(target~.,data=heart,kernel='polynomial')
> print(s)

Call:
svm(formula = target ~ ., data = heart, kernel = "polynomial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1
     degree:  3
     coef.0:  0

Number of Support Vectors:  182

> table(predict(s,heart),heart$target)

     0   1
  0 117   4
  1  21 161
```

오류율: 25/303

### 5-3-3. sigmoid커널

```
> s=svm(target~.,data=heart,kernel='sigmoid')
> print(s)
```

```
Call:
svm(formula = target ~ ., data = heart, kernel = "sigmoid")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
      coef.0:  0

Number of Support Vectors:  130

> table(predict(s,heart),heart$target)

     0   1
  0 100  19
  1  38 146
```

오류율: 57/303

SVM모델은 radial basis function 커널을 사용할 때, 가장 오류율이 낮고, 24/303의 오류율이 나오므로 92.08%의 정확률을 가진다.

## 5-4. k-NN모델

test case는 기존 heart데이터에서 1과 0이 비슷하게 나뉘는 구간인 165번째를 기준으로 50개의 데이터를 추출해서 사용한다. k=1부터 k=9까지 실행한 결과를 도출해본다.

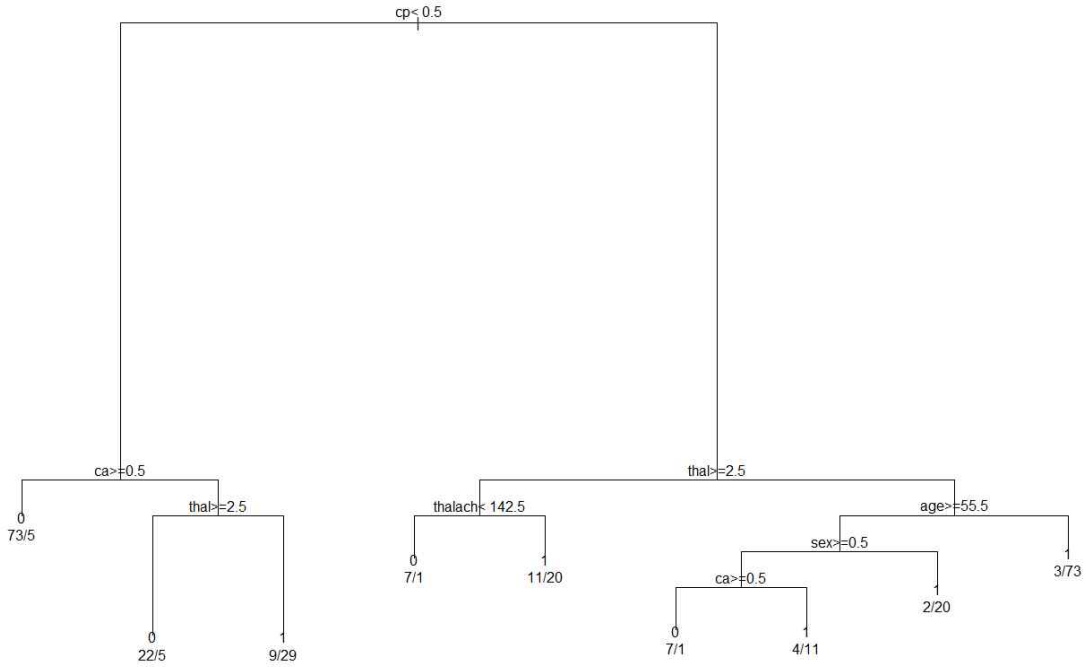
```
> test=heart[140:190,1:13]
> for(i in 1:9){
+ k=knn(train[,1:13],test,train$target,k=i)
+ prop.table(table(ifelse(k==heart[140:190,]$target,"True","False")))
+ }
True
   1


    False       True
0.1176471 0.8823529

    False       True
0.2156863 0.7843137

    False       True
0.2156863 0.7843137

    False       True
0.2156863 0.7843137

    False       True
0.2352941 0.7647059

    False       True
0.2352941 0.7647059

   False      True
0.254902 0.745098
```

```
    False      True
0.2941176 0.7058824
```

k=1일 때, 88.23%로 제일 높은 정확율을 보여준다.

## 5-5. 결정트리 모델

```
> library(rpart)
> r=rpart(target~.,data=heart)
> plot(r)
> text(r,use.n=TRUE)
```




```
> p=predict(r,heart,type='class')
> table(p,heart$target)
```

```
p       0     1
  0  109    12
  1   29  153
```

정확률= 262/303= 86.46%

# 5. 결론

각각의 모델에서 구한 정확률을 살펴보면 아래와 같다.

로지스틱 회귀 모델
52.32%

랜덤 포리스트 모델
82.28%

SVM 모델
92.08%

k-NN모델
88.23%

결정트리 모델
86.46%

Heart Disease UCI에 적합한 최상의 모델은 SVM모델(radial basis function 커널사용)입니다.