폐암
발병 가능성 예측

데이터사이언스학과
2020012021 표성주

Contents

폐암 발병 가능성 예측

01

문제 정의

폐암 발병 가능성 예측

- 2021년 통계청 자료 한국인 사망 원인 1위
- 2020년 보건복지부 자료 발생 2위 암종

24개 암종별	발생자수 (명)	상대빈도 (%)
모든 암(C00-C96)	247,952	100.0
갑상선(C73)	29,180	11.8
폐(C33-C34)	28,949	11.7
대장(C18-C20)	27,877	11.2
위(C16)	26,662	10.8
유방(C50)	24,923	10.1

(단위: 인구 10만 명당 명)

순위	사망원인	사망률	'20년 순위 대비
1	악성신생물(암)	161.1	-
2	심장 질환	61.5	-
3	폐렴	44.4	-
4	뇌혈관 질환	44.0	-
5	고의적 자해(자살)	26.0	-
6	당뇨병	17.5	-
7	알츠하이머병	15.6	-
8	간질환	13.9	-
9	패혈증	12.5	⬆(+1)
10	고혈압성 질환	12.1	⬇(-1)

02 # 데이터 수집

폐암 발병 가능성 예측

- 캐글에서 제공하는 폐암 발병 여부 데이터
- 성별, 연령, 흡연 여부, 가슴 통증 여부 등 총 16개 변수
- 총 309개의 데이터

```
> names(d)
 [1] "GENDER"                 "AGE"                    "SMOKING"
 [4] "YELLOW_FINGERS"         "ANXIETY"                "PEER_PRESSURE"
 [7] "CHRONIC.DISEASE"        "FATIGUE"                "ALLERGY"
[10] "WHEEZING"               "ALCOHOL.CONSUMING"      "COUGHING"
[13] "SHORTNESS.OF.BREATH"    "SWALLOWING.DIFFICULTY" "CHEST.PAIN"
[16] "LUNG_CANCER"
```

```
> ncol(d)
[1] 16
> nrow(d)
[1] 309
```

03 EDA

폐암 발병 가능성 예측

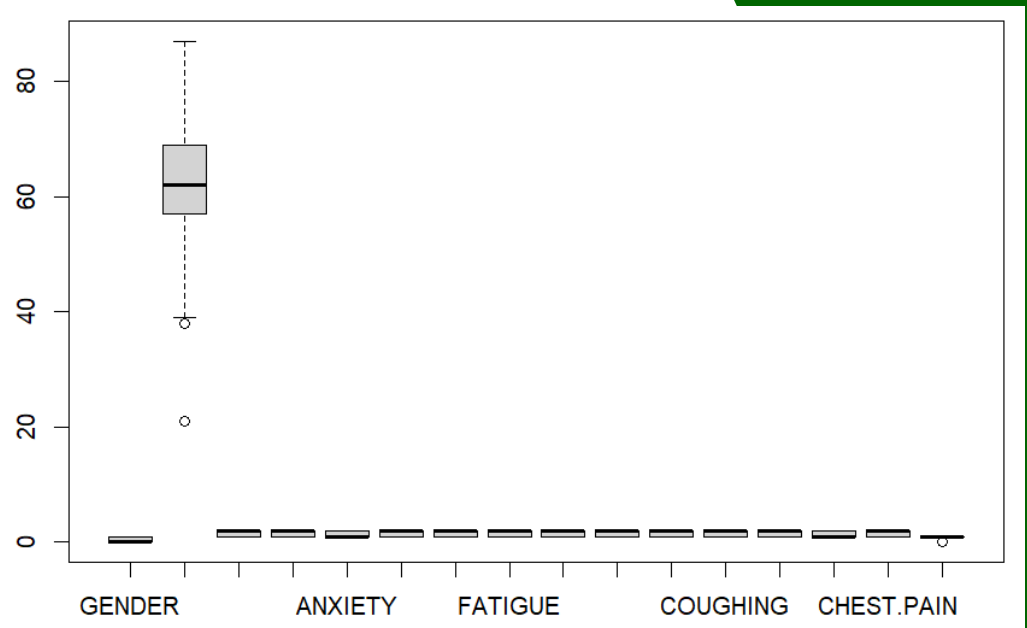
- 데이터 타입, 데이터 개수, 변수 개수 확인

```
> str(d)
'data.frame':   309 obs. of  16 variables:
 $ GENDER                : chr  "M" "M" "F" "M" ...
 $ AGE                   : int  69 74 59 63 63 75 52 51 68 53 ...
 $ SMOKING               : int  1 2 1 2 1 1 2 2 2 2 ...
 $ YELLOW_FINGERS        : int  2 1 1 2 2 2 1 2 1 2 ...
 $ ANXIETY               : int  2 1 1 2 1 1 1 2 2 2 ...
 $ PEER_PRESSURE         : int  1 1 2 1 1 1 1 2 1 2 ...
 $ CHRONIC.DISEASE       : int  1 2 1 1 1 2 1 1 1 2 ...
 $ FATIGUE               : int  2 2 2 1 1 2 2 2 2 1 ...
 $ ALLERGY               : int  1 2 1 1 1 2 1 2 1 2 ...
 $ WHEEZING              : int  2 1 2 1 2 2 2 1 1 1 ...
 $ ALCOHOL.CONSUMING     : int  2 1 1 2 1 1 2 1 1 2 ...
 $ COUGHING              : int  2 1 2 1 2 2 2 1 1 1 ...
 $ SHORTNESS.OF.BREATH   : int  2 2 2 1 2 2 2 2 1 1 ...
 $ SWALLOWING.DIFFICULTY: int  2 2 1 2 1 1 1 2 1 2 ...
 $ CHEST.PAIN            : int  2 2 2 2 1 1 2 1 1 2 ...
 $ LUNG_CANCER           : chr  "YES" "YES" "NO" "NO" ...
```

- 이상치 확인
- 결측치 확인

```
boxplot(d)
```

```
> is.null(d)
[1] FALSE
```



04

데이터 전처리

폐암 발병 가능성 예측

- 성별, 폐암 발병 여부 변수 수치형으로 변환
- 데이터의 개수를 맞추어 랜덤하게 샘플 추출

```
> d[d$LUNG_CANCER == "NO", "LUNG_CANCER"] = 0
> d[d$LUNG_CANCER == "YES", "LUNG_CANCER"] = 1
> d[d$GENDER == "M", "GENDER"] = 0
> d[d$GENDER == "F", "GENDER"] = 1
> d$GENDER = as.numeric(d$GENDER)
> d$LUNG_CANCER = as.numeric(d$LUNG_CANCER)
> str(d)
'data.frame':   309 obs. of  16 variables:
 $ GENDER                : num  0 0 1 0 1 1 0 1 1 0 ...
 $ AGE                   : int  69 74 59 63 63 75 52 51 68 53 ...
 $ SMOKING               : int  1 2 1 2 1 1 2 2 2 2 ...
 $ YELLOW_FINGERS        : int  2 1 1 2 2 2 1 2 1 2 ...
 $ ANXIETY               : int  2 1 1 2 1 1 1 2 2 2 ...
 $ PEER_PRESSURE         : int  1 1 2 1 1 1 1 2 1 2 ...
 $ CHRONIC.DISEASE       : int  1 2 1 1 1 2 1 1 1 2 ...
 $ FATIGUE               : int  2 2 2 1 1 2 2 2 2 1 ...
 $ ALLERGY               : int  1 2 1 1 1 2 1 2 1 2 ...
 $ WHEEZING              : int  2 1 2 1 2 2 2 1 1 1 ...
 $ ALCOHOL.CONSUMING     : int  2 1 1 2 1 1 2 1 1 2 ...
 $ COUGHING              : int  2 1 2 1 2 2 2 1 1 1 ...
 $ SHORTNESS.OF.BREATH   : int  2 2 2 1 2 2 2 2 1 1 ...
 $ SWALLOWING.DIFFICULTY: int  2 2 1 2 1 1 1 2 1 2 ...
 $ CHEST.PAIN            : int  2 2 2 2 1 1 2 1 1 2 ...
 $ LUNG_CANCER           : num  1 1 0 0 0 1 1 1 0 1 ...
```

```
library(dplyr)
library(PerformanceAnalytics)
library(magrittr)
```

```
> length(which(d$LUNG_CANCER==0))
[1] 39
> length(which(d$LUNG_CANCER==1))
[1] 270
```

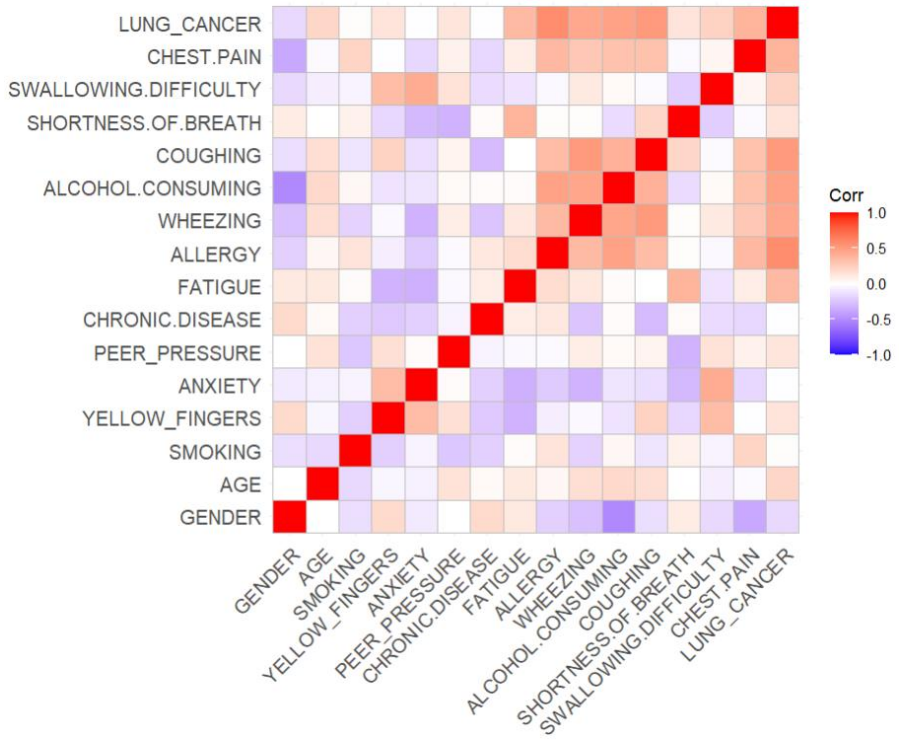
```
s = d[d$LUNG_CANCER==1,]
head(s)
s1 = sample_n(s, 40, replace = F)
s2 = d[d$LUNG_CANCER==0,]
S =rbind(s1, s2)
```

05 데이터 분석

폐암 발병 가능성 예측

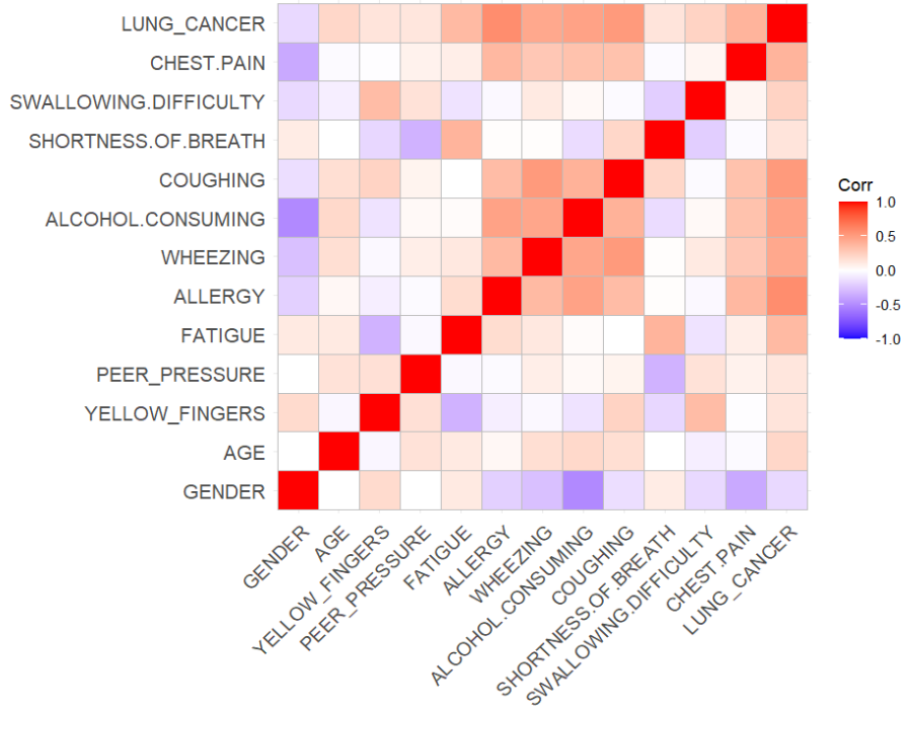
- 상관분석 진행

```
S_cor = cor(S)
round(S_cor, 2)
library(ggcorrplot)
ggcorrplot(S_cor) # 상관분석 그래프
```



- 종속변수와의 상관계수가 ± 0.01인 변수 제거 후 상관분석

```
D = subset(S, select=-c(SMOKING, ANXIETY, CHRONIC.DISEASE)) # 0.01 변수들 제거
round(cor(D), 2)
library(ggcorrplot)
ggcorrplot(cor(D))
```



05 데이터 분석

폐암 발병 가능성 예측

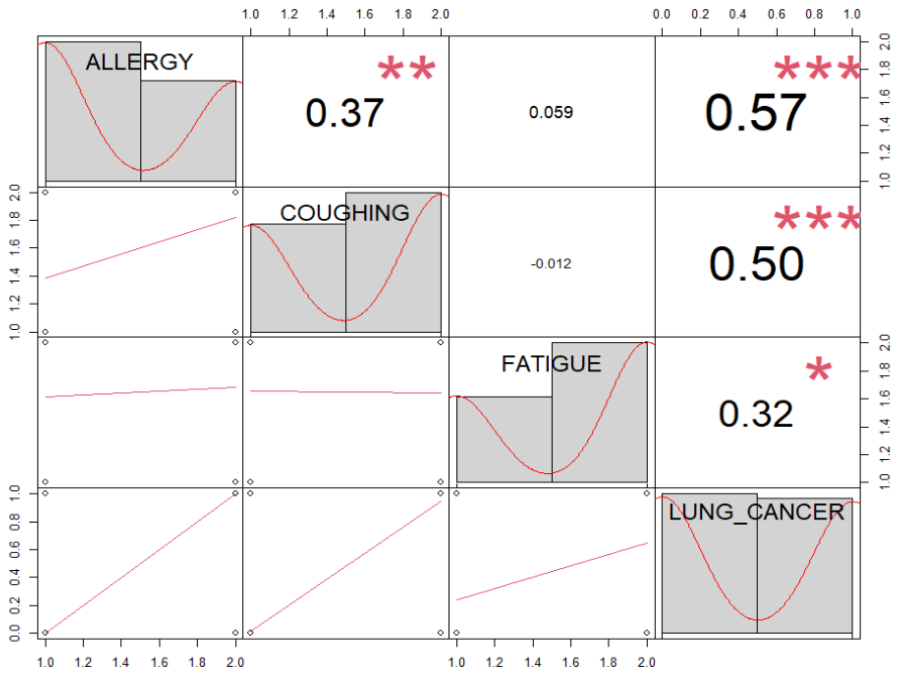
- train 데이터와 test 데이터 분리
- 변수 선택을 위해 stepwise selection 시행

```
train = sample_frac(D, size = 0.7)
test = sample_frac(D, size = 0.3)
```

```
#변수 선택
model = glm(LUNG_CANCER ~ 1, data = train)
ss = step(model, direction = "both", scope = LUNG_CANCER ~ GENDER + ALLERGY +
            YELLOW_FINGERS + SWALLOWING.DIFFICULTY + PEER_PRESSURE + FATIGUE +
            AGE + SHORTNESS.OF.BREATH + WHEEZING + ALCOHOL.CONSUMING + COUGHING +
            CHEST.PAIN)
> summary(ss)

Call:
glm(formula = LUNG_CANCER ~ ALLERGY + COUGHING + FATIGUE + SWALLOWING.DIFFICULTY +
    AGE, data = train)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.002848   0.367371  -5.452 1.62e-06 ***
ALLERGY              0.402881   0.099339   4.056 0.000179 ***
COUGHING             0.298677   0.101445   2.944 0.004939 **
FATIGUE              0.280527   0.095283   2.944 0.004940 **
SWALLOWING.DIFFICULTY 0.329582   0.117572   2.803 0.007229 **
AGE                  0.009755   0.005298   1.841 0.071608 .

SWALLOWING.DIFFICULTY   3.0774     1.7971   1.712 0.086817 .
```

- 로지스틱 회귀 모델 생성

```
# 최종모델
result = glm(LUNG_CANCER ~ ALLERGY + COUGHING + FATIGUE,
             data = train, family = binomial())
train %>% select_at(vars(ALLERGY, COUGHING,
      FATIGUE,  LUNG_CANCER)) %>% chart.Correlation(histogram = TRUE, pch=20)
```



06 # 결과 & 예측

폐암 발병 가능성 예측

- 모델 성능 확인
- train 데이터: 84%, test 데이터: 88%

```
> coef(result)
(Intercept)     ALLERGY    COUGHING     FATIGUE
 -13.726140    2.978732    2.782049    3.009662
```

```
> library(rsq)
> rsq(result)
[1] 0.5262261
```

```
Train = train[c("ALLERGY", "COUGHING", "FATIGUE", "LUNG_CANCER")]
Test = test[c("ALLERGY", "COUGHING", "FATIGUE", "LUNG_CANCER")]
SSS = D[c("ALLERGY", "COUGHING", "FATIGUE", "LUNG_CANCER")]
```

```
#오버/언더피팅 확인
p1 <- predict(result, newdata=Train, type="response")
round(p1)
table(round(p1),Train$LUNG_CANCER)
```

```
    0  1
 0 21  2
 1  7 25
```

```
p2 <- predict(result, newdata=Test, type="response")
round(p2)
table(round(p2),Test$LUNG_CANCER)
```

```
    0  1
 0 12  1
 1  2  9
```

06

결과 & 예측

폐암 발병 가능성 예측

- 알레르기 증상 여부, 기침 여부, 극심한 피로감 여부

- 알레르기 증상 여부: 1 = 없음 / 2 = 있음
- 기침 여부: 1 = 없음 / 2 = 있음
- 극심한 피로감 여부: 1 = 없음 / 2 = 있음
- 폐암 발병 여부: 0 = 발병하지 않음 / 1 = 발병

- 사람1,2는 폐암이 발병하지 않았고
  사람3,4는 폐암이 발병했을 것이라고 예측

```r
#예 측
man1 = data.frame(rbind(c(1,1,1)))
names(man1) = names(SSS)[1:3]
pred = predict(result, man1, type="response")
pred

man2 = data.frame(rbind(c(1,2,1)))
names(man2) = names(SSS)[1:3]
pred2 = predict(result, man2, type="response")
pred2

man3 = data.frame(rbind(c(2,1,2)))
names(man3) = names(SSS)[1:3]
pred3 = predict(result, man3, type="response")
pred3

man4 = data.frame(rbind(c(2,2,2)))
names(man4) = names(SSS)[1:3]
pred4 = predict(result, man4, type="response")
pred4
```

```
> pred
          1
0.006993908
> pred2
         1
0.1021419
> pred3
         1
0.7374384
> pred4
         1
0.9784321
```

THANK YOU