asthma dataset

Sai Pallavi

2024-09-20

question:1

```
library(tidyselect)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
`Asthma data set` <- read.csv("C:/Users/saipa/OneDrive/Desktop/dataset/dataset.csv")
```

```
#View(`Asthma data set`)
```

```
head(`Asthma data set`)
```

```
##   PatientID Age Gender Ethnicity EducationLevel      BMI Smoking
## 1      5034  63      0         1              0 15.84874       0
## 2      5035  26      1         2              2 22.75704       0
## 3      5036  57      0         2              1 18.39540       0
## 4      5037  40      1         2              1 38.51528       0
## 5      5038  61      0         0              3 19.28380       0
## 6      5039  21      0         2              0 21.81298       0
##   PhysicalActivity DietQuality SleepQuality PollutionExposure PollenExposure
## 1        0.8944483    5.488696     8.701003         7.3884806       2.855578
## 2        5.8973295    6.341014     5.153966         1.9698383       7.457665
## 3        6.7393670    9.196237     6.840647         1.4605930       1.448189
## 4        1.4045027    5.826532     4.253036         0.5819053       7.571845
## 5        4.6044926    3.127048     9.625799         0.9808746       3.049807
## 6        0.4700439    1.759118     9.549262         1.7114456       7.192424
##   DustExposure PetAllergy FamilyHistoryAsthma History.of.allergies. eczema
## 1    0.9743394          1                   1                     0      0
## 2    6.5846312          0                   0                     1      0
```

```
## 3    5.4457989          0                  1                    1      0
## 4    3.9653156          0                  0                    0      0
## 5    8.2606054          0                  0                    0      0
## 6    6.8320476          1                  0                    0      1
##   HayFever GastroesophagealReflux LungFunctionFEV1 LungFunctionFVC Wheezing
## 1        0                      0         1.369051        4.941206        0
## 2        0                      0         2.197767        1.702393        1
## 3        1                      0         1.698011        5.022553        1
## 4        1                      0         3.032037        2.300159        1
## 5        1                      0         3.470589        3.067944        1
## 6        0                      0         2.328191        5.898515        1
##   ShortnessOfBreath ChestTightness Coughing NighttimeSymptoms ExerciseInduced
## 1                 0              1        0                 0               1
## 2                 0              0        1                 1               1
## 3                 1              1        0                 1               1
## 4                 0              1        1                 1               0
## 5                 1              1        0                 0               1
## 6                 0              1        0                 0               1
##   Diagnosis DoctorInCharge
## 1         0      Dr_Confid
## 2         0      Dr_Confid
## 3         0      Dr_Confid
## 4         0      Dr_Confid
## 5         0      Dr_Confid
## 6         0      Dr_Confid
```

```
summary(`Asthma data set`$Gender)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4933  1.0000  1.0000
```

this as a dataset with nearly an equal number of Male and Female individuals, since the mean is approximately 0.4933.

```
summary(`Asthma data set`$Smoking)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1417  0.0000  1.0000
```

#Only about 14.17% of the individuals in your dataset are smokers, while the majority (85.83%) are non-smokers.

```
# Calculate probabilities for Gender
prob.gender <- prop.table(table(`Asthma data set`$Gender))
# Display probabilities
prob.gender
```

```
##
##        0        1
## 0.506689 0.493311
```

```r
# Calculate probabilities for Smoking
prob.Smoking <- prop.table(table(`Asthma data set`$Smoking))
# Display probabilities
prob.Smoking
```

```
##
##         0         1
## 0.8582776 0.1417224
```

```r
# Total population count
total.population <- nrow(`Asthma data set`)

# Display the total population count
total.population
```

```
## [1] 2392
```

```r
P.males=0.506689
p.females=0.493311
p.smoking.males=0.8582776
p.smoking.females=0.1417224
total.population=2392
```

```r
# Define parameters
p.male <- 0.506689
p.female <- 0.493311
p.Smoking.male <- 0.8582776
p.Smoking.female <- 0.1417224
total.population <- 2392

# Create empty vectors to store results
Gender <- vector("numeric", total.population)
Smoking <- vector("numeric", total.population)

# Set the seed for reproducibility
set.seed(2023)

# Assign gender (0 for male, 1 for female)
Gender <- sample(c(0, 1), size = total.population, prob = c(p.male, p.female),
                 replace = TRUE)

# Assign smoking status based on gender
for (k in 1:total.population) {
  if (Gender[k] == 0) {
    Smoking[k] <- sample(c(0, 1), prob = c(1 - p.Smoking.male, p.Smoking.male),
                         size = 1, replace = TRUE) # 0 for non-smoker, 1 for smoker
  }

  if (Gender[k] == 1) {
    Smoking[k] <- sample(c(0, 1), prob = c(1 - p.Smoking.female, p.Smoking.female),
                         size = 1, replace = TRUE)
  }
```

```
}

# View results
addmargins(table(Gender, Smoking))
```

```
##       Smoking
## Gender    0    1  Sum
##    0    174 1090 1264
##    1    989  139 1128
##    Sum 1163 1229 2392
```

```
#Probability of female smoking
sum(Gender == 1 & Smoking == 1)/total.population
```

```
## [1] 0.05811037
```

#5.81% are female smokers

```
#Probability of smoking
sum(Smoking)/total.population
```

```
## [1] 0.513796
```

#51.37% of the total population are smokers

question 2

```
contingency <- `Asthma data set` |>
  select(Smoking,Diagnosis)
contingency_table <- table(contingency$Smoking,contingency$Diagnosis)
contingency_table
```

```
##
##        0    1
##   0 1943  110
##   1  325   14
```

```
# Extracting values from the matrix
TP <- contingency_table["1", "1"]  # True Positives
TN <- contingency_table["0", "0"]    # True Negatives
FP <- contingency_table["1", "0"]   # False Positives
FN <- contingency_table["0", "1"]     # False Negatives

# Sensitivity
sensitivity <- TP / (TP + FN)
```

```r
# Specificity
specificity <- TN / (TN + FP)

# Prevalence
total_population <- sum(contingency_table)
prevalence <- (TP + FN) / total.population

# Display results
cat("Sensitivity:", sensitivity, "\n")
```

```
## Sensitivity: 0.1129032
```

```r
cat("Specificity:", specificity, "\n")
```

```
## Specificity: 0.8567019
```

```r
cat("Prevalence:", prevalence, "\n")
```

```
## Prevalence: 0.05183946
```

```r
# Parameters
prevalence <- 0.051          # Change this value as needed (e.g., 10%)
sensitivity <- 0.11        # Change to your sensitivity value
specificity <- 0.85        # Change to your specificity value
total.population <-2392     # Total population size

# Expected number of cases
expected.cases <- total.population * prevalence
cat("Expected Cases:", expected.cases, "\n")
```

```
## Expected Cases: 121.992
```

```r
# Expected number of non-cases
expected.noncases <- total.population - expected.cases
```

```r
expected.noncases = total.population - expected.cases
expected.noncases
```

```
## [1] 2270.008
```

```r
expected.true.positives = expected.cases * sensitivity
expected.true.positives
```

```
## [1] 13.41912
```

```r
expected.false.positives = expected.noncases * (1 - specificity)
expected.false.positives
```

```
## [1] 340.5012
```

```
total.expected.positives = expected.true.positives + expected.false.positives
total.expected.positives
```

```
## [1] 353.9203
```

```
expected.false.negatives = expected.cases * (1 - sensitivity)
expected.false.negatives
```

```
## [1] 108.5729
```

```
expected.true.negatives = expected.noncases * specificity
expected.true.negatives
```

```
## [1] 1929.507
```

```
total.expected.negatives = expected.true.negatives + expected.false.negatives
total.expected.negatives
```

```
## [1] 2038.08
```

```
ppv=expected.true.positives/total.expected.positives
npv = expected.true.negatives/total.expected.negatives
ppv
```

```
## [1] 0.03791565
```

```
npv
```

```
## [1] 0.9467279
```