KNN을 활용한 콜레스테롤 환자 분류

2131450 양 소 연

목 차

1. 연구의 배경

연구 제목	KNN을 활용한 콜레스테롤 환자 분류
연구 배경	식습관 변화로 인해 젊은 층에서의 고지혈증 발병률이 높아지고 있다. 고지혈증의 원인 중 하나는 '비만'이며 이것은 콜레스테롤 수치와 관련이 있다. 따라서 연령과 비만율에 따른 콜레스테롤 환자를 진단해보고자 한다.
연구 문제	콜레스테롤 지수가 높은 환자를 진단하는 KNN 모델 생성

1. 연구의 배경

기존 연구

- 일반적으로 고지혈증은 중년층(50~60대 중후반)에서 발생함
- 연령이 높아짐에 따라 체중이 증가하므로 혈액 지방 함량이 높아짐

기존 연구와의 차이점

- 고지혈증이 발생하는 원인이 나이가 아닌 다른 변수인 것으로 가정하여 고지혈증 발생 원인 조사
- 혈액 지방 함량에 연령과 체중 중 어떤 변수의 영향이 큰지 분석
- 특성에 따라 고지혈증 환자 예측

"30~40대 젊은층에서 혈관막는 '고지혈증' 증가 추세"

주로 중년 이후에게 많은 것으로 알려진 고지혈증이 최근 서구화된 식생활과 잦은 음주, 스트레스, 운동부족 등으로 **30~40대 젊은층에서도 환자가 상당수인 것으로 나타났다.**

국민건강보험공단에 따르면 고지혈증으로 병원을 찾은 사람은 2008년 74만6000명에서 2013년 128만8000명으로 크게 늘어 매년 11.5%가량 증가률을 보였다. 그 중 40대의 경우는 2008년 14만명에 그쳤던 것이 2015년 24만명으로 7년새 10만명이 늘어나며 70%가량 급증한 것으로 집계됐다.

이보람, "30~40대 젊은층에서 혈관막는 '고지혈증' 증가 추세", 헬스조선, 2016. 03. 24,
https://m.health.chosun.com/article/article.html?contid=2016032400849#a

2. 데이터 연구 방법론

콜레스테롤 데이터

데이터 출처	https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
데이터 설명	13 columns / 70,000 row
	Age / Height / Weight / Gender / Systolic blood pressure / Diastolic blood pressure / Cholesterol / Glucose / Smoking / Alcohol intake / Physical activity / Presence or absence of cardiovascular disease
데이터 특이사항	Age는 나이가 아닌 days가 단위 Height(cm), Weight(kg) 주요변수인 Cholesterol은 1, 2, 3으로 분류

```
head(data)
   id    age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio
 0 18393      2    168     62   110    80           1    1     0    0      1      0
 1 20228      1    156     85   140    90           3    1     0    0      1      1
 2 18857      1    165     64   130    70           3    1     0    0      0      1
 3 17623      2    169     82   150   100           1    1     0    0      1      1
 4 17474      1    156     56   100    60           1    1     0    0      0      0
 8 21914      1    151     67   120    80           2    2     0    0      0      0
```

About Dataset

Data description

There are 3 types of input features:

- *Objective*: factual information;
- *Examination*: results of medical examination;
- *Subjective*: information given by the patient.

Features:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All of the dataset values were collected at the moment of medical examination.

2. 데이터 연구 방법론

1. 이상치 처리

- Age(day) 를 연령대 별로 나누어 줌

```r
# age / 365 -> 단위가 day 임
data_new <- data %>% mutate(year = age/365)
data_new <- data_new %>% relocate(year, .after = age)

# 연령별로 묶어주기
data_new$year <- floor(data_new$year)
data_new <- data_new %>% mutate(age_range = ifelse(year < 30, 20,
                                                   ifelse(year < 40, 30,
                                                          ifelse(year<50, 40,
                                                                 ifelse(year<60, 50,
                                                                        ifelse(year<70, 60))))))
```

- Height의 이상치 제거

```r
# height의 이상치 제거
library(psych)
descr <- describe(data$height)
descr <- descr %>% mutate(LL = mean - 2*sd)
descr <- descr %>% mutate(UL = mean + 2*sd)

table(data_new$height > descr$UL)
data_new <- data_new %>% filter(height <= descr$UL)
```

```
      age             year      
 Min.   :10798   Min.   :29.58  
 1st Qu.:17664   1st Qu.:48.39  
 Median :19703   Median :53.98  
 Mean   :19469   Mean   :53.34  
 3rd Qu.:21327   3rd Qu.:58.43  
 Max.   :23713   Max.   :64.97  
```

- Weight의 이상치 제거

```r
# weight의 이상치 제거
descr <- describe(data$weight)
descr <- descr %>% mutate(LL = mean - 2*sd)
descr <- descr %>% mutate(UL = mean + 2*sd)

table(data_new$weight > descr$UL)
data_new <- data_new %>% filter(weight <= descr$UL)
```

- BMI(비만율) 변수 추가

```r
# BMI 변수 생성: 체질량지수는 자신의 몸무게(kg)를 키의 제곱(m)으로 나눈 값입니다.
data_new <- data_new %>% mutate(bmi = round(weight/(height/100)^2, 2))
```

2. 데이터 연구 방법론

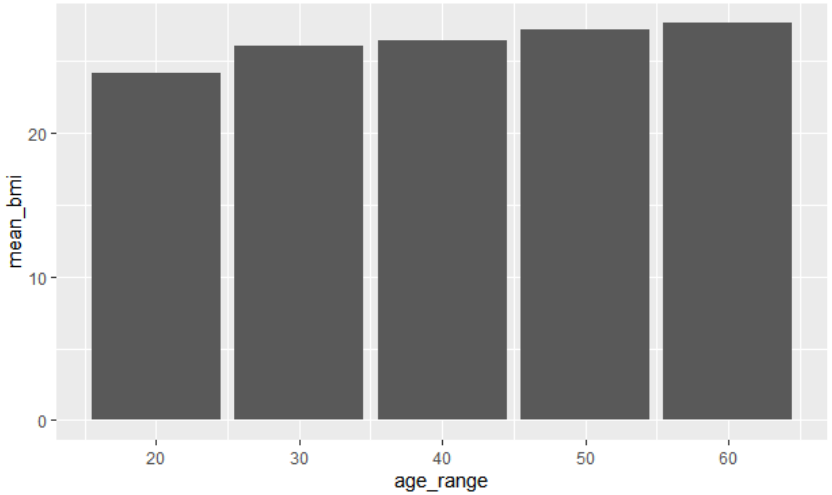

```
head(data_new)
id    age year age_range gender height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio   bmi
 0 18393   50        50      2    168     62   110    80           1    1     0    0      1      0 21.97
 1 20228   55        50      1    156     85   140    90           3    1     0    0      1      1 34.93
 2 18857   51        50      1    165     64   130    70           3    1     0    0      0      1 23.51
 3 17623   48        40      2    169     82   150   100           1    1     0    0      1      1 28.71
 4 17474   47        40      1    156     56   100    60           1    1     0    0      0      0 23.01
 8 21914   60        60      1    151     67   120    80           2    2     0    0      0      0 29.38
```

2. height와 weight의 이상치 제거

```
     height          weight      
Min.   : 55.0   Min.   : 10.00  
1st Qu.:159.0   1st Qu.: 65.00  
Median :165.0   Median : 72.00  
Mean   :164.4   Mean   : 74.21  
3rd Qu.:170.0   3rd Qu.: 82.00  
Max.   :250.0   Max.   :200.00  
```

```
     height          weight      
Min.   : 55.0   Min.   : 10.00  
1st Qu.:159.0   1st Qu.: 64.00  
Median :165.0   Median : 71.00  
Mean   :164.2   Mean   : 72.56  
3rd Qu.:169.0   3rd Qu.: 80.00  
Max.   :250.0   Max.   :102.00  
```

```
age_range `n()` `mean(bmi, na.rm = T)` `sd(bmi, na.rm = T)`
    <dbl> <int>                   <dbl>                <dbl>
       20     3                    24.2                 5.44
       30  1667                    26.1                 6.00
       40 18439                    26.4                 4.73
       50 33552                    27.2                 5.17
       60 12354                    27.6                 5.43
```



연령대가 높아질수록 비만율이 증가함

2. 데이터 연구 방법론

3. train 데이터셋과 test 데이터셋 구성

```
> table(df_z$cholesterol == "2") / 5000

FALSE  TRUE
0.754 0.246
> table(df_train$cholesterol == "2") / 3513

    FALSE      TRUE
0.7577569 0.2422431
> table(df_test$cholesterol == "2") / 1487

    FALSE      TRUE
0.7451244 0.2548756
```

4. 다섯가지 변수와 콜레스테롤 변수와의 중요도 비교

```
> varImp(knn.train, scale = F)
ROC curve variable importance

       Importance
bmi        0.6110
weight     0.5939
age        0.5734
height     0.5334
gender     0.5209
```

5. train 데이터 프레임을 이용한 최적의 k값과 모델 도출

```
> control <- trainControl(method = "repeatedcv", number = 10, repeats = 10)
> set.seed(1234)
> knn.train <- train(cholesterol~., data = df_train, method = "knn", trControl = control, tuneGrid = grid1)
> knn.train ### K = 12
k-Nearest Neighbors

3513 samples
   5 predictor
   2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 3162, 3161, 3161, 3162, 3161, 3162, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   3  0.7095965  0.08991657
   4  0.7101099  0.07693484
   5  0.7245413  0.07402157
   6  0.7249103  0.07020367
   7  0.7296644  0.05796122
   8  0.7292920  0.05315944
   9  0.7350415  0.04869130
  10  0.7346162  0.04708014
  11  0.7400238  0.05387882
  12  0.7404516  0.05096520

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 12.
```

7만개의 데이터를 R에서 처리하지 못하여 5,000개 추출

콜레스테롤 수치는 나이보다 비만율의 영향을 받는다.

2. 데이터 연구 방법론

6. Test 데이터 프레임을 이용한 Knn.train 성능 평가

```
> confusionMatrix(pred.test1, df_test$cholesterol)
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 1059  345
         2   49   34

               Accuracy : 0.735
                 95% CI : (0.7118, 0.7573)
    No Information Rate : 0.7451
    P-Value [Acc > NIR] : 0.822

                  Kappa : 0.0612

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.95578
            Specificity : 0.08971
         Pos Pred Value : 0.75427
         Neg Pred Value : 0.40964
             Prevalence : 0.74512
         Detection Rate : 0.71217
   Detection Prevalence : 0.94418
      Balanced Accuracy : 0.52274

       'Positive' Class : 1
```

7. 성능 개선

```
> confusionMatrix(pred.test2, df_test$cholesterol)
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 986 316
         2 122  63

               Accuracy : 0.7054
                 95% CI : (0.6816, 0.7285)
    No Information Rate : 0.7451
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.0675

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8899
            Specificity : 0.1662
         Pos Pred Value : 0.7573
         Neg Pred Value : 0.3405
             Prevalence : 0.7451
         Detection Rate : 0.6631
   Detection Prevalence : 0.8756
      Balanced Accuracy : 0.5281

       'Positive' Class : 1
```

성능이 저하되었다 → 더 많은 변수를 추가하여 모델 생성

2. 데이터 연구 방법론

8. 콜레스테롤 변수와 나머지 변수들 간의 중요도 비교

```
ROC curve variable importance

       Importance
gluc       0.6541
ap_hi      0.6304
cardio     0.6199
bmi        0.6110
weight     0.5939
ap_lo      0.5919
age        0.5734
height     0.5334
gender     0.5209
active     0.5176
smoke      0.5098
alco       0.5069
```

콜레스테롤은 혈당과 관련이 있으며
나이보다는 비만율의 영향을 받는다.

9. train 데이터 프레임을 이용한 최적의 k값과 모델 도출

```
k-Nearest Neighbors

3513 samples
  12 predictor
   2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 3162, 3161, 3161, 3162, 3161, 3162, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   3  0.7503524  0.2368311
   4  0.7478474  0.2284730
   5  0.7649264  0.2482843
   6  0.7666359  0.2483434
   7  0.7756021  0.2596899
   8  0.7740947  0.2519899
   9  0.7779652  0.2566524
  10  0.7795322  0.2618426
  11  0.7812398  0.2623090
  12  0.7808700  0.2575063

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 11.
```

2. 데이터 연구 방법론

10. 변수를 추가한 Test 데이터를 이용하여 Knn.train 성능 평가

```
> confusionMatrix(pred.test1, df_test$cholesterol)
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 1042  276
         2   66  103

               Accuracy : 0.77
                 95% CI : (0.7478, 0.7912)
    No Information Rate : 0.7451
    P-Value [Acc > NIR] : 0.01422

                  Kappa : 0.2595

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9404
            Specificity : 0.2718
         Pos Pred Value : 0.7906
         Neg Pred Value : 0.6095
             Prevalence : 0.7451
         Detection Rate : 0.7007
   Detection Prevalence : 0.8863
      Balanced Accuracy : 0.6061

       'Positive' Class : 1
```

11. 변수를 추가한 Knn.train 모델의 성능 개선

```
> confusionMatrix(pred.test2, df_test$cholesterol)
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 1068  287
         2   40   92

               Accuracy : 0.7801
                 95% CI : (0.7582, 0.8009)
    No Information Rate : 0.7451
    P-Value [Acc > NIR] : 0.0009395

                  Kappa : 0.263

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9639
            Specificity : 0.2427
         Pos Pred Value : 0.7882
         Neg Pred Value : 0.6970
             Prevalence : 0.7451
         Detection Rate : 0.7182
   Detection Prevalence : 0.9112
      Balanced Accuracy : 0.6033

       'Positive' Class : 1
```

변수 추가 후, 0.735 → 0.7801 정확도 개선

3. 연구결과

12. Predict 데이터를 생성하여 KNN 모델로 예측

- 원본 데이터에서 사용하지 않은 데이터 중 100개를 추출하여 predict data 생성

- train, test 데이터와 같은 방식으로 데이터 전처리

- Predict data에서 cholesterol 열을 제거하여 예측 진행

13. 12개의 변수를 활용하여 생성한 kknn 모델을 사용하여 콜레스테롤 환자 예측하기

```
# 정답 2121211122
# knn 2111211111 7개
# kknn 2111211111 7개
```

```
df_pred <- as.data.frame(scale(df_pred))
pred.test3 <- predict(knn.train, newdata = df_pred)
pred.test3 #  1 1 1 2 1 1 1 1 1
pred.test4 <- predict(kknn.train, newdata = df_pred)
pred.test4 # 2 1 1 1 2 1 1 1 1 1
```

약 70%의 확률로 콜레스테롤 환자를 예측하는 모델 생성

4. 결론

- 콜레스테롤 수치는 단순히 나이가 아닌 비만율의 영향을 많이 받는다.
- 젊은 층에서 고지혈증 환자 발생 비율이 높아지는 것은 식습관 변화로 인하여 비만율이 높아지기 때문이다.
- 올바른 식습관 개선으로 성인병 위험율을 낮출 수 있을 것으로 기대된다.
- 연령별로 해당 데이터를 살펴보기에는 20,30대의 데이터가 부족하였으므로, 추후 더 많은 데이터 확보가 필요하다.
- 콜레스테롤 진단의 예측 정확도를 높이기 위하여 모델의 적합성에 대한 연구가 필요하다.

5. 참고문헌

Data Reference :

- Kilpatrick, D. & Cameron-Jones, M. (1998). Numeric prediction using instance-based learning with encoding length selection. In Progress in Connectionist-Based Information Systems. Singapore: Springer-Verlag.

참고문헌 :

- Johns Hopkins Medicine, https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel
- 이보람, "30~40대 젊은층에서 혈관막는 '고지혈증' 증가 추세", 헬스조선, 2016. 03. 24, https://m.health.chosun.com/article/article.html?contid=2016032400849#a

감사합니다