재현가능한 데이터 사이언스(Data Science) - BMI

Reproducible Data Science

2019-08-29

# 비즈니스 설명

[캐글, “500 Person Gender-Height-Weight-Body Mass Index - Height and Weight random generated, Body Mass Index Calculated”](https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex)에서 데이터를 바탕으로 고객에게 체중과 키 정보만 제공하면 체질량 지수(Body Mass Index , BMI)를 예측하는 모형을 개발하여 고객이 궁금해하는 서비스를 개발하고자 한다.

**체질량 지수(體質量指數, Body Mass Index, BMI)**는 인간의 비만도를 나타내는 지수로, 체중과 키의 관계로 계산된다. 키가 t 미터, 체중이 w 킬로그램일 때, BMI는 다음이 수식으로 표현된다. (키의 단위가 센티미터가 아닌 미터임에 유의해야 한다.)

체질량지수 (BMI지수)로 과체중 혹은 비만을 판정하는 한국 사례 [[1]](#footnote-22)

|  |  |
| --- | --- |
| 구분 | BMI 지수 |
| 고도 비만 | 40 이상 |
| 중등도 비만 (2단계 비만) | 35 - 39.9 |
| 경도 비만 (1단계 비만) | 30 - 34.9 |
| 과체중 | 25 - 29.9 |
| 정상 | 18.5 - 24.9 |
| 저체중 | 18.5 미만 |

# 데이터

[캐글, “500 Person Gender-Height-Weight-Body Mass Index - Height and Weight random generated, Body Mass Index Calculated”](https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex)에서 데이터를 바탕으로 고객에게 체중과 키 및 라벨 데이터 **index**가 준비되어 있어 키와 몸무게를 통해 BMI 예측한다.

## 데이터 사전

* Gender : Male / Female
* Height : Number (cm)
* Weight : Number (Kg)
* Index :
  + 0 : Extremely Weak
  + 1 : Weak
  + 2 : Normal
  + 3 : Overweight
  + 4 : Obesity
  + 5 : Extreme Obesity

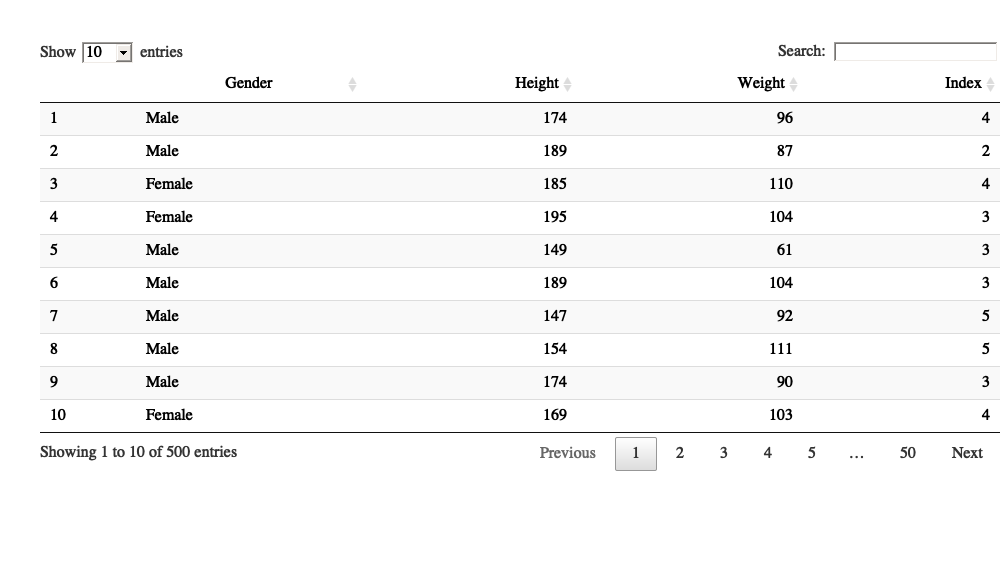
## 데이터 가져오기

캐글에서 내려받은 원본 데이터를 살펴본다.

library(tidyverse)  
library(here)  
  
file\_path <- here("data", "500\_Person\_Gender\_Height\_Weight\_Index.csv")  
  
bmi\_dat <- read\_csv(file\_path)  
  
glimpse(bmi\_dat)

Observations: 500  
Variables: 4  
$ Gender <chr> "Male", "Male", "Female", "Female", "Male", "Male", "Male…  
$ Height <dbl> 174, 189, 185, 195, 149, 189, 147, 154, 174, 169, 195, 15…  
$ Weight <dbl> 96, 87, 110, 104, 61, 104, 92, 111, 90, 103, 81, 80, 101,…  
$ Index <dbl> 4, 2, 4, 3, 3, 3, 5, 5, 3, 4, 2, 4, 3, 2, 2, 5, 5, 5, 5, …

bmi\_dat %>%   
 DT::datatable()



# 탐색적 데이터 분석

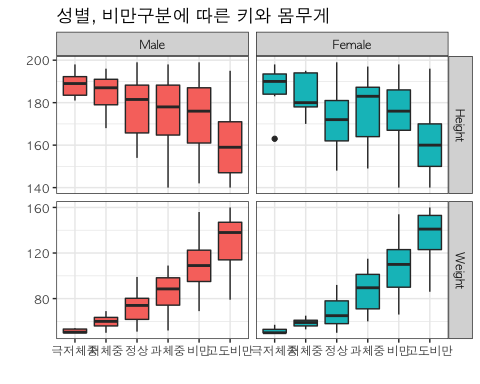
## 요약 통계

bmi\_df <- bmi\_dat %>%   
 mutate(Index = factor(Index, levels = c(0,1,2,3,4,5), labels = c("극저체중", "저체중", "정상", "과체중", "비만", "고도비만")),  
 Gender = factor(Gender, levels = c("Male", "Female")))  
  
bmi\_df %>%   
 group\_by(Index) %>%   
 summarise(평균키 = mean(Height),  
 평균체중 = mean(Weight))

# A tibble: 6 x 3  
 Index 평균키 평균체중  
 <fct> <dbl> <dbl>  
1 극저체중 188. 51.7  
2 저체중 185. 59.4  
3 정상 174. 69.1  
4 과체중 176. 86.9  
5 비만 174. 108.   
6 고도비만 161. 133.

## 시각화

bmi\_df %>%   
 gather(키체중, 값, -Gender, -Index) %>%   
 ggplot(aes(x=Index, y= 값, fill=Gender)) +  
 geom\_boxplot(show.legend = FALSE) +  
 facet\_grid(키체중 ~ Gender, scales="free") +  
 theme\_bw(base\_family = "AppleGothic") +  
 labs(x="", y="",  
 title="성별, 비만구분에 따른 키와 몸무게")



# 예측모형 - BMI 예측

BMI 그룹: “극저체중”, “저체중”, “정상”, “과체중”, “비만”, “고도비만”

# 0. 환경설정 ------  
library(caret)  
library(doSNOW)  
  
set.seed(777)  
  
# 1. 데이터 ------  
# bmi\_df  
  
# 2. 데이터 전처리 ------  
  
# 3. 예측모형 ------  
## 3.1. 병렬처리 환경설정  
num\_cores <- parallel:::detectCores()  
start\_time <- Sys.time()  
  
cl <- makeCluster(num\_cores, type = "SOCK")  
registerDoSNOW(cl)  
  
## 3.2. 훈련 vs 검증/시험  
train\_test\_index <- createDataPartition(bmi\_df$Index, p = 0.7, list = FALSE)  
  
train <- bmi\_df[train\_test\_index, ]  
test <- bmi\_df[-train\_test\_index, ]  
  
## 3.3. 모형 개발/검증 데이터셋 준비 ------  
cv\_folds <- createMultiFolds(train$Index, k = 10, times = 5)  
cv\_ctrl <- trainControl(method = "cv", number = 10,  
 index = cv\_folds,   
 verboseIter = TRUE)  
## 3.2. 예측모형 적용  
### ranger  
gc\_ranger\_model <- train(Index ~., train,  
 method = "ranger",  
 tuneLength = 7,  
 trControl = cv\_ctrl)

note: only 2 unique complexity parameters in default grid. Truncating the grid to 2 .  
  
Aggregating results  
Selecting tuning parameters  
Fitting mtry = 2, splitrule = extratrees, min.node.size = 1 on full training set

# 4. 모형 비교평가-----  
gc\_pred\_class <- predict(gc\_ranger\_model, newdata = test, type="raw")  
## 혼동행렬 -----  
bmi\_conf <- confusionMatrix(gc\_pred\_class, test$Index)  
  
bmi\_conf$table

Reference  
Prediction 극저체중 저체중 정상 과체중 비만 고도비만  
 극저체중 3 0 0 0 0 0  
 저체중 0 4 1 0 0 0  
 정상 0 2 15 1 0 0  
 과체중 0 0 2 16 0 0  
 비만 0 0 2 3 38 4  
 고도비만 0 0 0 0 1 55

cat("정확도: ", scales::percent(bmi\_conf$overall[["Accuracy"]]))

정확도: 89.1%

stopCluster(cl)

# BMI 예측

bmi\_test\_dat <- tribble(  
 ~"Gender", ~"Height", ~"Weight",  
 "Male", 149, 61,  
 "Female", 172, 67  
)  
  
predict(gc\_ranger\_model, newdata = bmi\_test\_dat, type="raw")

[1] 과체중 정상   
Levels: 극저체중 저체중 정상 과체중 비만 고도비만

1. [위키백과, “체질량 지수” (2019-04-22 접근함)](https://ko.wikipedia.org/wiki/%EC%B2%B4%EC%A7%88%EB%9F%89_%EC%A7%80%EC%88%98) [↑](#footnote-ref-22)