한국인 당뇨병 주요 변수 예측을 위한 통계분석 및 머신러닝 적용

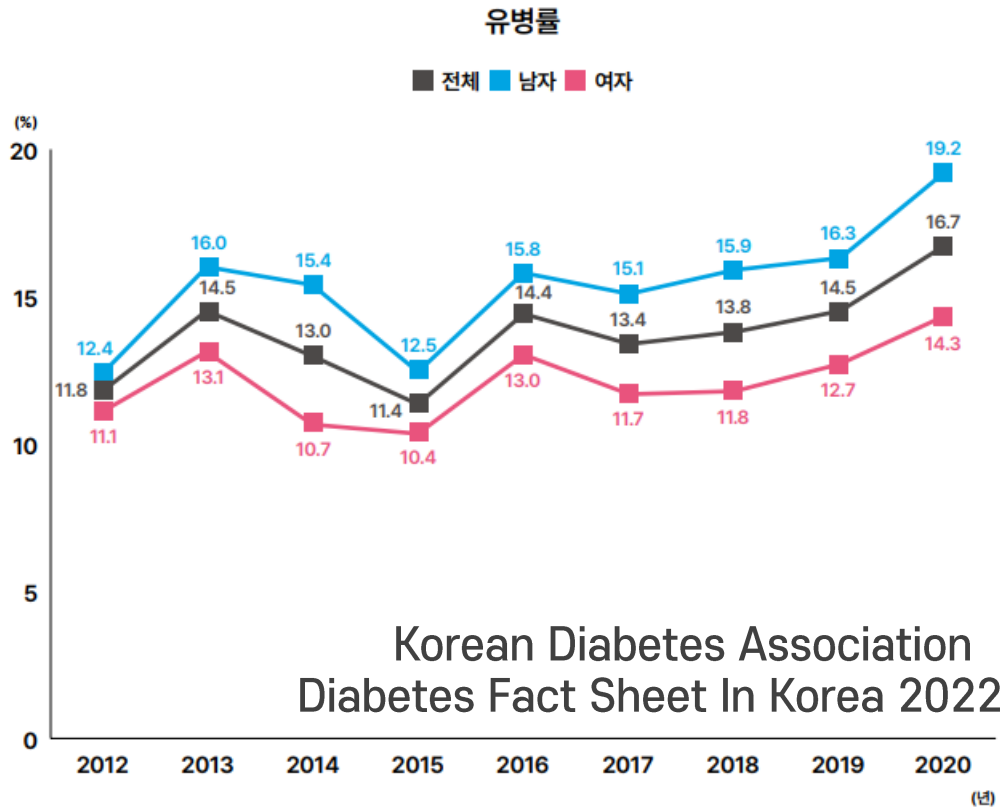
1조 | 슈가걸

김유빈 김윤정 김정희 서주희 임수아 전원희

>> Table of contents

Part 1 >> 주제 선정 배경

최근 9년간 당뇨병 유병률 및 인구 변화



Korean Diabetes Association
Diabetes Fact Sheet In Korea 2022

30세 이상 성인 7명 중 1명 당뇨병 환자

조정희 기자 | 입력 2020.10.15 14:59 | 댓글 0

"지구촌 당뇨 환자, 2050년 13억명으로 2배로"

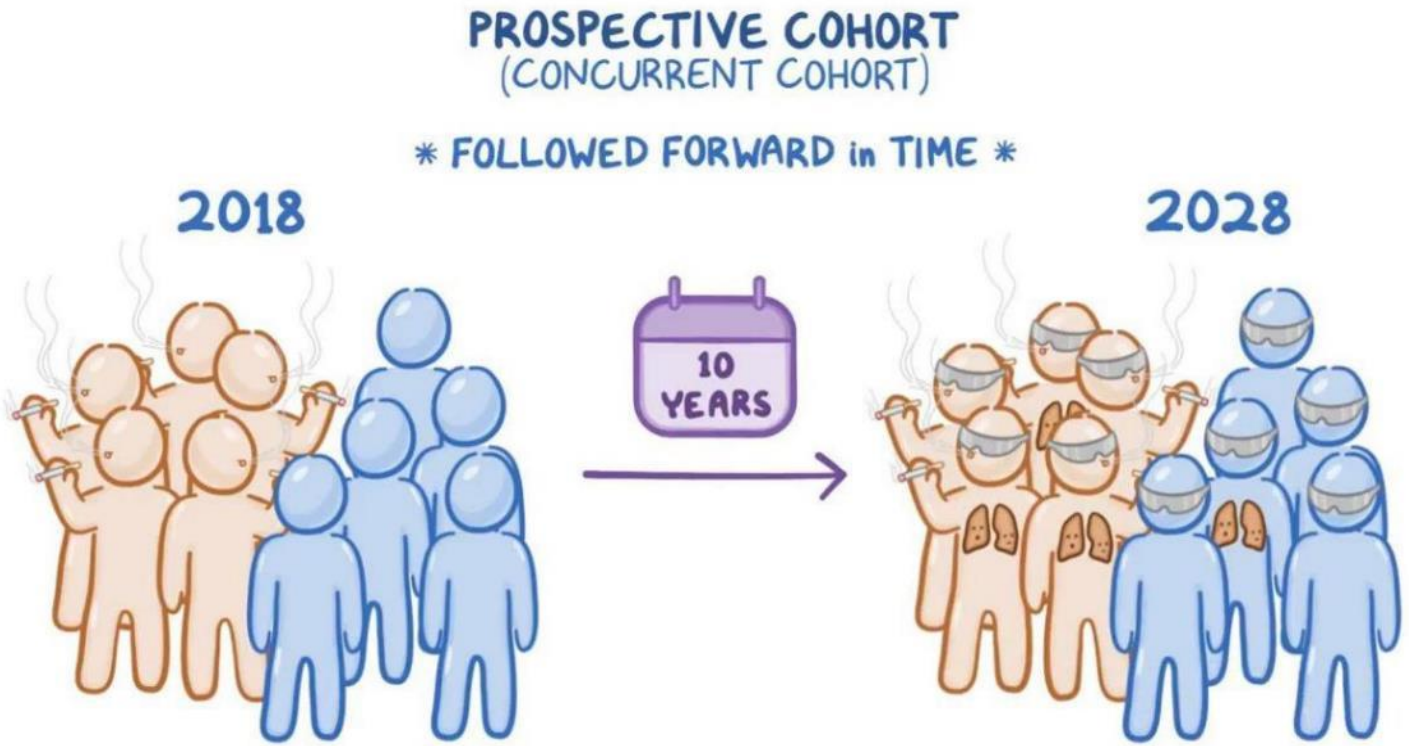
김도산 기자 | 입력 2023.06.23 16:24 | 수정 2023.06.23 21:41 | 댓글 0

<2018-2021 당뇨병 환자 수>

(출처 : 국가건강정보)

한국뿐만 아니라 전세계적으로 증가하는 당뇨병 환자

Part 1 >> 주제 선정 배경 - 데이터셋 소개 (KoGES)



한국인유전체역학조사사업(KoGES)
한국인에게 흔한 만성 질환의 위험요인을 규명하여 맞춤·예방의학 구현의 과학적 근거를 마련하고자 질병관리청 국립보건연구원이 수행하는 코호트 사업

Part 1 >> 주제 선정 배경 - 데이터셋 소개

안성안산코호트 기반 역학정보 공개데이터 → **총 2년 간격, 8번의 추적 조사 데이터 (1~8기)**

안성안산코호트 기반(1기) 역학정보 공개데이터 통합코드북(

변수 내용							데이터현황	
공개 여부	테이블명 (국문)	테이블명 (영문)	변수명	변수설명	변수값(코드) 설명 (66666=조사안함, 77777=해당 없음, 99999=미상/무응답/미	변수유형	안성	안산
공개	질병과거력	AS1_06_DISEASE	AS1_PdDm	당뇨병 진단받은 경험유무	1=아니오, 2=예	범주형	●	●
공개	질병과거력	AS1_06_DISEASE	AS1_PdDmAg	당뇨병 진단시 나이	만( )세	연속형	●	●
공개	질병과거력	AS1_06_DISEASE	AS1_PdUl	위염/위궤양 진단받은 경험유무	1=아니오, 2=예	범주형	●	●
공개	질병과거력	AS1_06_DISEASE	AS1_PdUlAg	위염/위궤양 진단시 나이	만( )세	연속형	●	●
공개	질병과거력	AS1_06_DISEASE	AS1_PdAl	알레르기질환 진단받은 경험유무	1=아니오, 2=예	범주형	●	●
공개	질병과거력	AS1_06_DISEASE	AS1_PdAlAg	알레르기질환 진단시 나이	만( )세	연속형	●	●
공개	질병과거력	AS1_06_DISEASE	AS1_PdMi	심근경색 진단받은 경험유무	1=아니오, 2=예	범주형	●	●

변수 유형
: 연속형 & 범주형

77777 : 해당없음
99999 : 무응답

✓ 1기 데이터만 사용한 이유

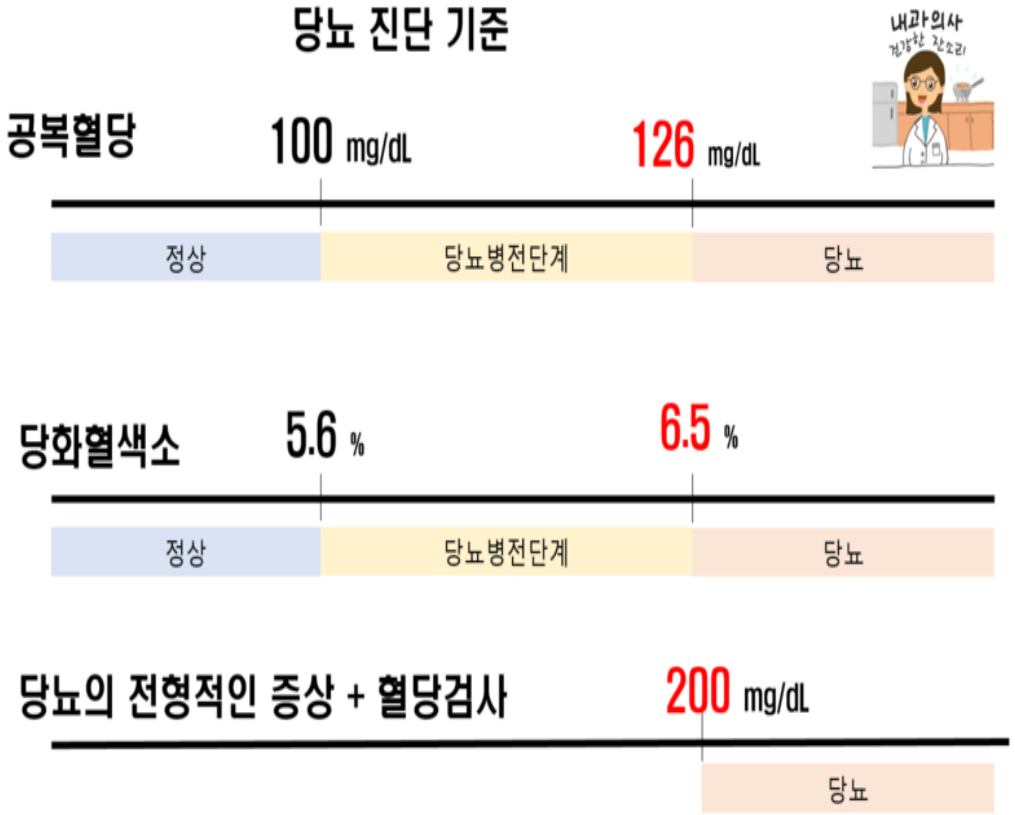
추적 조사 과정에서

연구대상자의 조사 유무 변화가 큼

➔ 예비 연구로 1기 데이터를 사용해 예측하는 것이 가장 정확하다고 판단

Part 1 >> 주제 선정 배경 - 당뇨병 환자 진단 기준

당뇨병 환자 진단 기준 : 당화혈색소 & 공복 혈당



당화혈색소 6.5 이상 or
공복혈당 126 이상: 당뇨병 O

그 외 : 당뇨병 X

➔ 가장 정확한 결과값 도출

Part 2 >> # 데이터 전처리

1938개 변수 ➔ 52개 변수

피마 인디언 당뇨병 자료를 이용한
다중 로지스틱 회귀모형 비교

이 왕 선

강원대학교대학원 통계학과

당뇨병은 인슐린의 분비량이 부족하거나 정상적인 기능이 이루어지지 않는 대사질환의 일종이다. 당뇨병은 신장 질환, 신경 손상, 실명, 혈관 손상의 발병 위험을 증가시키며, 이는 심장 질환에 기여한다. 그렇기 때문에 당뇨병 진단의 분류는 중요한 문제이다.

당뇨병 진단을 예측하는 가장 일반적인 방법은 로지스틱 회귀모형이다. 로지스틱 회귀모형은 해석이 쉽고 선형 회귀모형과의 유사성으로 인해 사용이 쉽다는 장점을 가지고 있다. 당뇨병과 소화기 및 신장 질환 국립 연구소에서 조사한 피마 인디언 당뇨병 자료를 이용하여 여러 가지 로지스틱 회귀모형들을 비교한다. 또한 로지스틱 회귀모형 분석 결과 나이가 유의하지 않은 이유에 대해서 설명한다. 각 모형에서의 성능의 차이는 오분류율, AUROC, 민감도, 특이도로 비교한다.

불필요한 하위 변수 제외

✓ 기존 변수의 경우 상위 단계 변수와 하위변수 모두 포함

(ex. 음주 기간, 금주 기간, 막걸리 음주 여부, 맥주 음주 1회 주량 등)

✓ 이상치, 결측치 포함

➔ **대표 상위 단계 변수만 남김!**

당뇨병 진단 지표 대표적 8가지

나이 / 흡연여부 / 음주량 / 허리둘레 / 혈압약 복용여부 / 생활습관 /
총 콜레스테롤 / 인슐린 등 임상 지표

➔ **생활습관 + 임상 지표**

➢ 강원대학교 대학원의 "피마 인디언 당뇨병 자료를 이용한 다중 로지스틱 회귀모형 비교" 논문 참조 하여 변수 추출

Part 2 >> 데이터 전처리 ➡ 남자, 여자 데이터 따로 추출

임상수치, bmi, 허리둘레 등 남,여의 기준치가 다름

➔ 예측모델을 따로 만들어야 정확해짐

EX)

엉덩이 둘레 평균 측정 기준치

성인 남성

평균 범위: 94-102 cm (37-40 **인치**)

성인 여성

평균 범위: 90-98 cm (35-38 **인치**)

BMI 측정 기준치

남성

- 저체중: BMI 18.4 미만
- 정상 체중: BMI 18.5에서 22.9 사이
- 과체중: BMI 23에서 24.9 사이
- 비만: BMI 25 이상

여성

- 저체중: BMI 18.4 미만
- 정상 체중: BMI 18.5에서 23.9 사이
- 과체중: BMI 24에서 24.9 사이
- 비만: BMI 25 이상

복부 지방율 측정 기준치

성인 남성

낮음: **복부 지방율** 6-12%

보통: **복부 지방율** 13-20%

높음: **복부 지방율** 21-25%

매우 높음: **복부 지방율** 26% **이상**

성인 여성

낮음: **복부 지방율** 16-22%

보통: **복부 지방율** 23-29%

높음: **복부 지방율** 30-35%

매우 높음: **복부 지방율** 36% **이상**

Part 2 >>

데이터 전처리

연속형 데이터 ➔ 범주형 데이터로 변환

- ✓ 기존에 수치로 작성 되어 있는 데이터 값을 범위로 쉽게 구분 ➔ **해석성 향상**
- ✓ 범주로 변환 시 **이상치의 영향을 줄이는데 도움을** 준다고 판단
- ✓ 임상 지표의 경우 **알려진 정보를 기준으로** 범주화

EX)

대한진단검사의학회
Korean Society for Laboratory Medicine

(1) 공복상태 혈당 검사 (fasting glucose)

[정상치] 70~99 ㎎/ℓ 8시간 이상 공복 후 측정한 혈당이 126 mg/dL 이상인 경우 당뇨병으로 진단이 됩니다. 당뇨병의 증상이 없다면 한번 더 측정한 후 두 번의 결과를 보고 판정을 내리는 것이 정확합니다. 공복혈당이 100-125 mg/dl 사이로 나온다면 이것도 정상이 아니고 공복혈당장애(impaired fasting glucose)로 분류합니다. 이는 당뇨병 전단계 또는 당뇨병이 생길 위험도가 높은 상태인데 그 위험도는 공복혈당장애가 있는 사람이 1년이 지나면 약 10%에서 당뇨병이 생긴다고 합니다.

(2) 75g 경구 당부하 검사

포도당 75g을 녹인 용액을 마시고 2시간 후 측정한 혈당이 200 mg/dL 이상인 경우 당뇨병으로 진단이 됩니다. 포도당을 마신 후에는 가만히 앉아 있다가 측정하는 것이 좋습니다.

* 정맥혈로 혈당검사를 하는 경우의 당뇨병 진단기준*

세브란스 | 건강정보

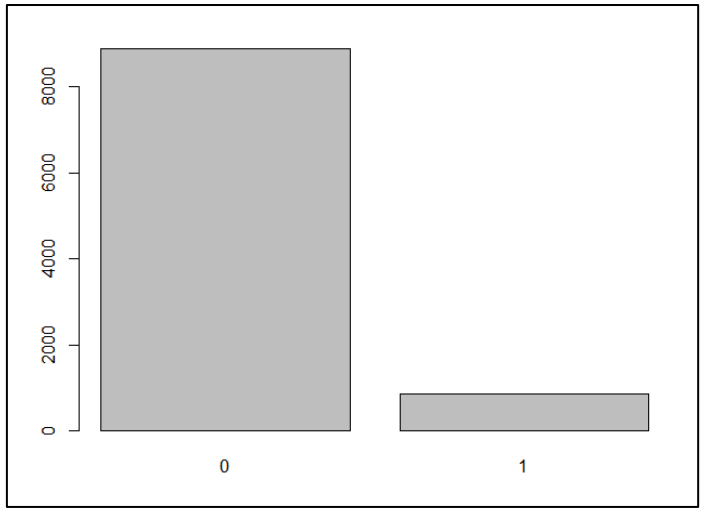
당화혈색소 (%)	평균 혈장 혈당 (mg/dL)
5.7 이하	100 미만
6	126
7	154
8	183
9	212
10	240
11	269
12	300

Part 2 >> 데이터 전처리

```
f1<- read.csv(file = "변수.csv")
install.packages("dplyr")
library(dplyr)

f2<-subset(f1, !AS1_GLUO_ORI==99999) #9731명
f3<-subset(f2, !AS1_HBA1C==99999) #9730명


# 당뇨병 변수(dm) 생성 ----
f3 <- f3 %>% mutate(dm = ifelse(AS1_HBA1C >=  6.5 | AS1_GLUO_ORI >= 126, 1,
                                ifelse(AS1_HBA1C <  6.5 & AS1_GLUO_ORI < 126,0, NA)))
table(f1$dm)

# 빈도 분석 ----
install.packages("descr")
library(descr)
table.dm <- freq(f3$dm)
round(table.dm, digits = 2)

freq(f3$dm)
```

당화혈색소 6.5 이상 or
공복혈당 126 이상: 당뇨병 O

→ **DM 변수 생성**

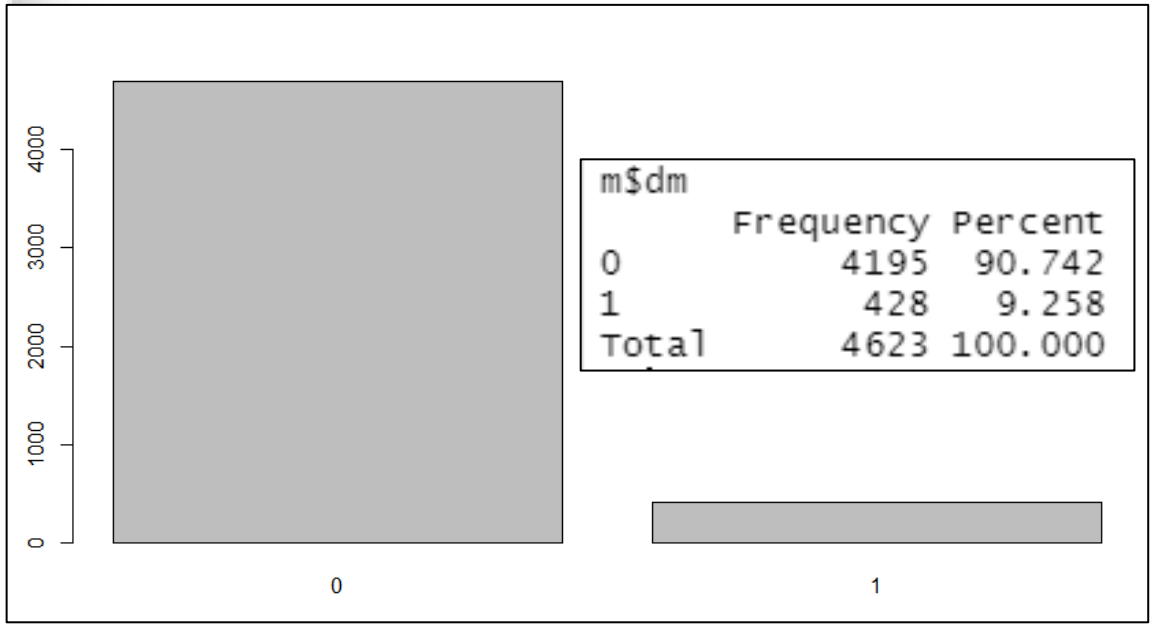



```
f3$dm
      Frequency Percent
0          8884   91.31
1           846    8.69
Total      9730  100.00
```

Part 2 >> 데이터 전처리

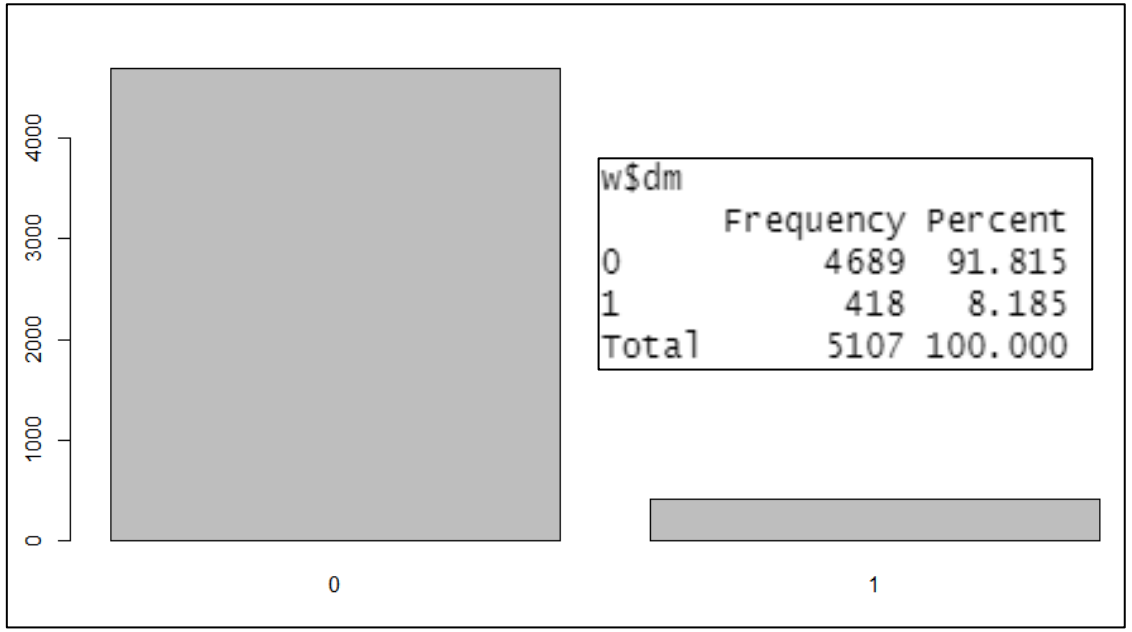
```
m<- subset(f3, AS1_SEX == 1) #4623명
w<- subset(f3, AS1_SEX == 2) #5107명

freq(m$dm)
freq(w$dm)
```

성별 별 당뇨유무



m$dm	Frequency	Percent
0	4195	90.742
1	428	9.258
Total	4623	100.000

남



w$dm	Frequency	Percent
0	4689	91.815
1	418	8.185
Total	5107	100.000

여

Part 2 >> 데이터 전처리

```
f3 [f3==99999]<-NA

f4 <- f3 %>% select(-AS1_FMDMREL1A, -AS1_FMHTREL1A)
```

➡ 가족력 변수 2개 삭제

```
f4 [f4==99999]<-NA
f4_1<-na.omit(f4)
```

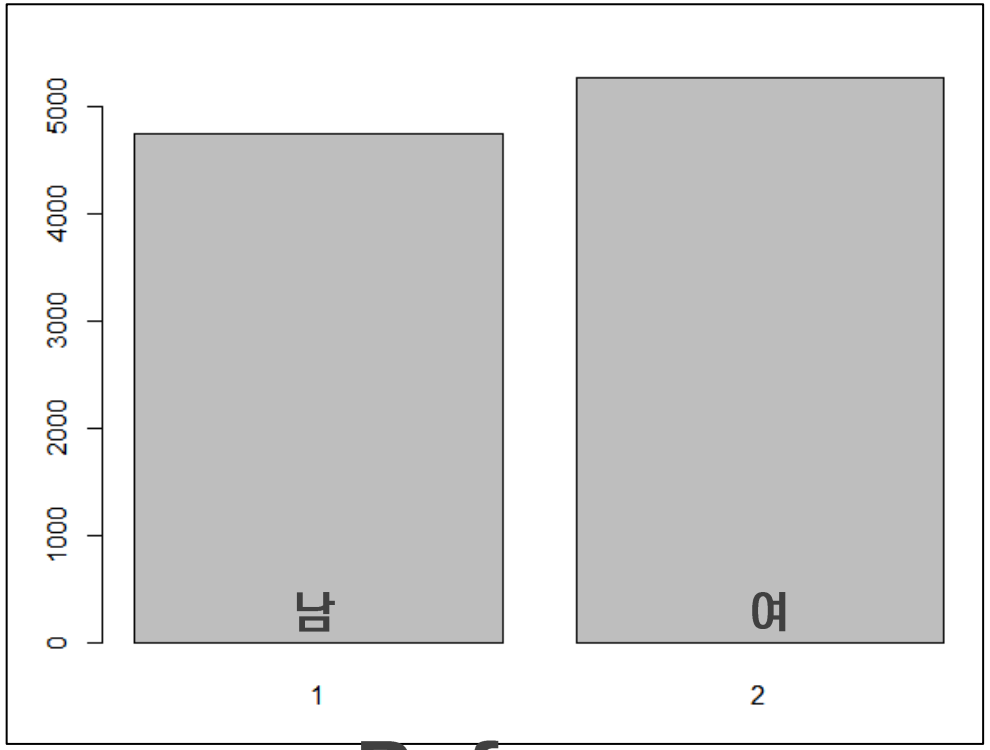
➡ 결측치 및 이상치 제거

```
# 수정된 데이터 프레임을 새로운 CSV 파일로 저장
write.csv(f4, file = "가족력지움.csv", row.names = FALSE)

freq(f4_1$AS1_SEX) # W :48%, M:52%
```

Part 2 >>

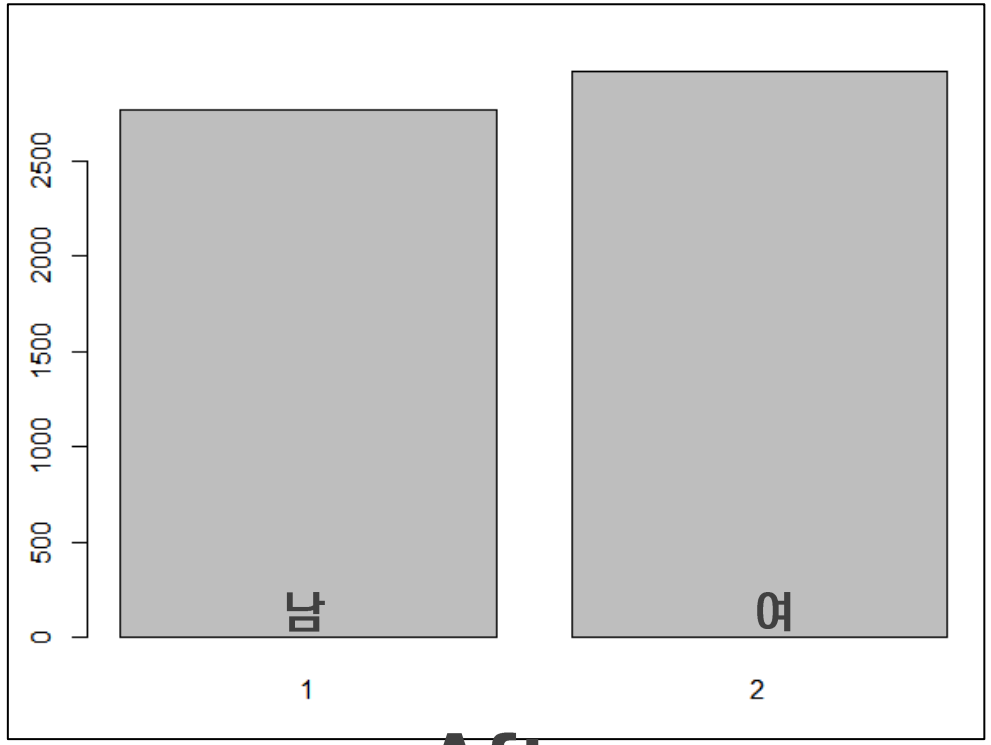
데이터 전처리

```
> freq(f1$AS1_SEX)  #W: 47%, M: 52%
f1$AS1_SEX
      Frequency Percent
1          4758   47.44
2          5272   52.56
Total     10030  100.00
```



Before

```
> freq(f4_1$AS1_SEX) # W :48%, M:51%
f4_1$AS1_SEX
      Frequency Percent
1          2769   48.28
2          2966   51.72
Total      5735  100.00
```



After

Part 2 >> 데이터 전처리

```
install.packages("readr")
library(readr)

# 엑셀 파일 경로 지정
csv_file_path <- "가족력지움.csv"

# 엑셀 파일 읽기
data <- read_csv(csv_file_path)

# '77777' 값을 0으로 대체
data[data == 77777] <- 0

# 변경된 데이터를 새로운 엑셀 파일로 저장
write.csv(data, "가족력에서7을0으로.csv", row.names = FALSE) #54

##########################################################################
setwd("C:/Users/82108/Desktop")

library(dplyr)
cleaned_data <- f4_1 %>% filter(!is.na(AS1_BMI))

n1<- read.csv(file = "bmi로2066날림.csv")
dim(n1)


# 변경된 데이터 프레임 저장
write.csv(n1, file = "final.csv", row.names = FALSE)
```

결측치 및 이상치 제거

결측치 및 이상치 제거

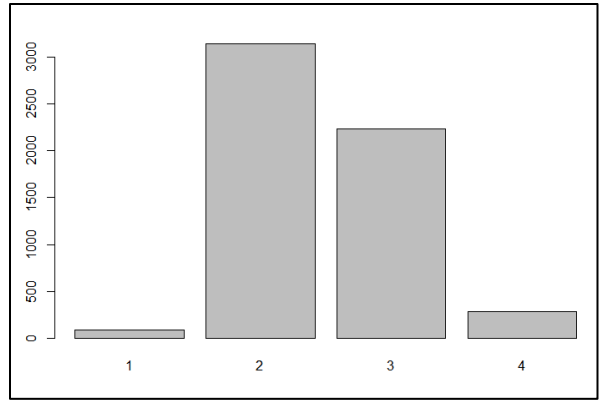
Part 2 >> 데이터 전처리

```r
#BMI
n1$BMI_CATEGORY <- ifelse(n1$AS1_BMI < 18.5, "1",
                          ifelse(n1$AS1_BMI >= 18.5 & n1$AS1_BMI <= 24.9, "2",
                                 ifelse(n1$AS1_BMI >= 25 & n1$AS1_BMI <= 29.9, "3", "4")))

freq(n1$BMI_CATEGORY)

#HIP1
n1$HIP1_CATEGORY <- ifelse(n1$AS1_SEX==1 & n1$AS1_HIP1 >= 94 & n1$AS1_HIP1 <= 102, "1", "2")

n1$HIP1_CATEGORY <- ifelse(n1$AS1_SEX==2 & n1$AS1_HIP1 >= 90 & n1$AS1_HIP1 <= 98, "1","2")

freq(n1$HIP1_CATEGORY)


#BDCMSC
n1$BDCMSC_CATEGORY <- ifelse(n1$AS1_SEX==1 & n1$AS1_BDCMSC >=35 & n1$AS1_BDCMSC <= 42, "1",

                             ifelse(n1$AS1_SEX==1 & n1$AS1_BDCMSC >= 43 & n1$AS1_BDCMSC <= 52, "2", "3"))

n1$BDCMSC_CATEGORY <- ifelse(n1$AS1_SEX==2 & n1$AS1_BDCMSC >=27 & n1$AS1_BDCMSC <= 35, "1",
                             ifelse(n1$AS1_SEX==2 & n1$AS1_BDCMSC >= 36 & n1$AS1_BDCMSC <= 45, "2", "3"))

freq(n1$BDCMSC_CATEGORY)
```

* 임상 지표가 성별에 따라 다른 경우

```r
#ABFTR
n1$ ABFTR_CATEGORY <- ifelse(n1$AS1_SEX==1 & n1$AS1_ABFTR < 0.9, "1","2")


n1$ ABFTR_CATEGORY <- ifelse(n1$AS1_SEX==2 &n1$AS1_ABFTR < 0.85, "1", "2")
```

```
n1$BMI_CATEGORY
      Frequency Percent
1            85   1.482
2          3136  54.682
3          2230  38.884
4           284   4.952
Total      5735 100.000
```



데이터 범주화

생활습관 ->Y/N 여부에 따라

임상 지표 ->알려진 기준에 따라

+ 성별에 따라 다르게 범주화!

연속형 변수 ➔ 범주형 변수
15개 변경

Part 3 >> 통계적 분석

```python
import pandas as pd
from scipy.stats import chi2_contingency

# 엑셀 파일을 읽어 데이터프레임으로 변환
file_path = 'C:/Users/82108/Desktop/변경된데이터2.csv'
df = pd.read_csv(file_path)

# AS1_SEX 변수 : 1은 남성, 2는 여성으로 변환
df['AS1_SEX'] = df['AS1_SEX'].replace({1: '남성', 2: '여성'})

# 변수 리스트
variables_list = [
    'AS1_AGE', 'AS1_RAMY_1', 'AS1_PIZA_1', 'AS1_CAND_1', 'AS1_COKE_1',
    'AS1_FRFSWT', 'AS1_FRFSLT', 'AS1_DRDUA', 'AS1_SMOKEA', 'AS1_PDHT',
    'AS1_PDDM', 'AS1_PDLP', 'AS1_PHYACTL', 'AS1_PHYACTM', 'AS1_PHYACTH',
    'AS1_MARRYA', 'AS1_JOBB', 'AS1_EDUA', 'AS1_INCOME', 'AS1_WAFQINC',
    'AS1_URINFQINC', 'AS1_TIED', 'AS1_WGTINC', 'AS1_TOOTH', 'AS1_INSM',
    'AS1_SLPAMTM', 'AS1_SNRTRT', 'AS1_DRINK', 'AS1_STRA', 'AS1_STRE',
    'AS1_STRG', 'AS1_STRC', 'AS1_STRD', 'GLUO_ORI_CATEGORY',
    'HDL_ORI_CATEGORY', 'TG_ORI_CATEGORY', 'PACKYR_CATEGORY',
    'HBA1C_CATEGORY', 'BDFTR_CATEGORY', 'BMI_CATEGORY', 'HIP1_CATEGORY',
    'BDCMSC_CATEGORY', 'BDCFT_□CATEGORY', 'ABFTR_CATEGORY', 'TOTALC_CATEGORY',
    'SLPAMTM_CATEGORY', 'BdFtR_CATEGORY', 'WC_CATEGORY', 'TCHL_CATEGORY',
    'INSO_CATEGORY']
```

```
AS1_AGE 변수에 대한 남성 다중 교차표:
dm          0    1
AS1_AGE
1        1500  133
2         591   76
3         400   69
남성 p-value: 6.820623701372336e-05
```

```
AS1_AGE 변수에 대한 여성 다중 교차표:
dm          0   1
AS1_AGE
1        1480  73
2         668  69
3         591  85
여성 p-value: 1.4117369065204385e-10
```

```
남성의 DM 비율 (%):
dm                0          1
AS1_AGE
1         60.216780  47.841727
2         23.725411  27.338129
3         16.057808  24.820144
```

```
여성의 DM 비율 (%):
dm                0          1
AS1_AGE
1         54.034319  32.158590
2         24.388463  30.396476
3         21.577218  37.444934
```

```python
# DM을 타겟으로 한 남성과 여성 각각에 대한 다중 교차표, p-value, 성별별 백분율
for var in variables_list:
    male_df = df[df['AS1_SEX'] == '남성']
    female_df = df[df['AS1_SEX'] == '여성']

    male_cross_tab = pd.crosstab(male_df[var], male_df['dm'])
    female_cross_tab = pd.crosstab(female_df[var], female_df['dm'])

    print(f"{var} 변수에 대한 남성 다중 교차표:")
    print(male_cross_tab)

    chi2_male, p_value_male, _, _ = chi2_contingency(male_cross_tab)
    print(f"남성 p-value: {p_value_male}")

    print(f"{var} 변수에 대한 여성 다중 교차표:")
    print(female_cross_tab)

    chi2_female, p_value_female, _, _ = chi2_contingency(female_cross_tab)
    print(f"여성 p-value: {p_value_female}")
    print(f"{var} 변수에 대한 여성 p-value: {p_value_female}")
    # 성별별 백분율 계산 및 출력
    male_percentage = male_cross_tab / male_cross_tab.sum() * 100
    female_percentage = female_cross_tab / female_cross_tab.sum() * 100

    print(f"남성의 DM 비율 (%):")
    print(male_percentage)

    print(f"여성의 DM 비율 (%):")
    print(female_percentage)
```

다중 교차표 & p-value & 성별과 변수에 따른 DM 분포

Part 3 >> 통계적 분석

```python
import pandas as pd
from scipy.stats import chi2_contingency

# 엑셀 파일을 읽어 데이터프레임으로 변환
file_path = 'C:/Users/82108/Desktop/변경된데이터2.csv'
df = pd.read_csv(file_path)

# 다중 교차표 생성 및 p-value 계산
def calculate_chi_square(df, variables):
    result_female = []
    result_male = []

    for var in variables:
        # 여성 그룹
        cross_tab_female = pd.crosstab([df[var], df['AS1_SEX']], df['dm'])
        cross_tab_female = cross_tab_female.xs(2, level='AS1_SEX')

        # 남성 그룹
        cross_tab_male = pd.crosstab([df[var], df['AS1_SEX']], df['dm'])
        cross_tab_male = cross_tab_male.xs(1, level='AS1_SEX')

        chi2_female, p_value_female, _, _ = chi2_contingency(cross_tab_female)
        chi2_male, p_value_male, _, _ = chi2_contingency(cross_tab_male)

        result_female.append((var, p_value_female))
        result_male.append((var, p_value_male))

    result_female.sort(key=lambda x: x[1])  # 여성 그룹 p-value 낮은 순으로 정렬
    result_male.sort(key=lambda x: x[1])    # 남성 그룹 p-value 낮은 순으로 정렬

    return result_female, result_male
```

```python
# 변수 리스트
variables_list = [
    'AS1_AGE', 'AS1_RAMV_1', 'AS1_PIZA_1', 'AS1_CAND_1', 'AS1_COKE_1',
    'AS1_FRFSWT', 'AS1_FRFSLT', 'AS1_DRDUA', 'AS1_SMOKEA', 'AS1_PDHT',
    'AS1_PDDM', 'AS1_PDLP', 'AS1_PHYACTL', 'AS1_PHYACTM', 'AS1_PHYACTH',
    'AS1_MARRYA', 'AS1_JOBB', 'AS1_EDUA', 'AS1_INCOME', 'AS1_WAFQINC',
    'AS1_URINFQINC', 'AS1_TIED', 'AS1_WGTINC', 'AS1_TOOTH', 'AS1_INSM',
    'AS1_SLPAMTM', 'AS1_SNRTRT', 'AS1_DRINK', 'AS1_STRA', 'AS1_STRE',
    'AS1_STRG', 'AS1_STRC', 'AS1_STRD', 'GLUO_ORI_CATEGORY',
    'HDL_ORI_CATEGORY', 'TG_ORI_CATEGORY', 'PACKYR_CATEGORY',
    'HBA1C_CATEGORY', 'BDFTR_CATEGORY', 'BMI_CATEGORY', 'HIP1_CATEGORY',
    'BDCMSC_CATEGORY', 'BDCFT_CATEGORY', 'ABFTR_CATEGORY', 'TOTALC_CATEGORY',
    'SLPAMTM_CATEGORY', 'BdFtR_CATEGORY', 'WC_CATEGORY', 'TCHL_CATEGORY',
    'INSO_CATEGORY']
```

```python
female_results, male_results = calculate_chi_square(df, variables_list)

# 여성 그룹 결과 출력
print("여성 그룹:")
for item in female_results:
    print(f"변수: {item[0]}, p-value: {item[1]}")

# 남성 그룹 결과 출력
print("남성 그룹:")
for item in male_results:
    print(f"변수: {item[0]}, p-value: {item[1]}")
```

여성 그룹:

```
TG_ORI_CATEGORY, p-value: 1.500938782454174e-24
TCHL_CATEGORY, p-value: 6.924321786736264e-20
WC_CATEGORY, p-value: 4.7851742939734795e-14
BMI_CATEGORY, p-value: 4.490211372503533e-12
AS1_PDHT, p-value: 1.1553353178259535e-11
AS1_AGE, p-value: 1.4117369065204385e-10
HDL_ORI_CATEGORY, p-value: 9.741948795230998e-06
AS1_MARRYA, p-value: 3.6437159299014955e-05
AS1_DRINK, p-value: 0.00010655728034577835
BdFtR_CATEGORY, p-value: 0.00027373451925549584
BDCMSC_CATEGORY, p-value: 0.0003866120997581154 4
AS1_EDUA, p-value: 0.0006761592676120409
AS1_WAFQINC, p-value: 0.0008799943944831905
BDCFT_CATEGORY, p-value: 0.0090010560297517 2
AS1_PDLP, p-value: 0.009144988329077189
AS1_PIZA_1, p-value: 0.010040832166293442
AS1_STRC, p-value: 0.03263024185948941
HIP1_CATEGORY, p-value: 0.048696772275056824
```

남성 그룹:

```
BDFTR_CATEGORY, p-value: 7.863147127607071e-11
TG_ORI_CATEGORY, p-value: 3.074590384819529e-09
BMI_CATEGORY, p-value: 1.2593663311956065e-07
TCHL_CATEGORY, p-value: 1.0403732219677528e-05
AS1_AGE, p-value: 6.820623701372336e-05
AS1_PDHT, p-value: 0.00011754657403788003
AS1_JOBB, p-value: 0.0006294710936010769
AS1_WAFQINC, p-value: 0.0009975312662764615
HDL_ORI_CATEGORY, p-value: 0.0011093721016843402
AS1_DRDUA, p-value: 0.0019815741321103485
AS1_INCOME, p-value: 0.007581261277940903
AS1_URINFQINC, p-value: 0.027342460795830175
AS1_WGTINC, p-value: 0.027821100499450073
AS1_TOOTH, p-value: 0.029654716974668854
AS1_SLPAMTM, p-value: 0.041180266547897644
```

유의미한 P-VALUE

Part 3 >> 통계적 분석 - Stepwise 통계 분석

남자

변수	p-value
음주기간	0.00399
고혈압 진단 여부	0.011257
당뇨병 진단 여부	0.017507
고지혈증 진단 여부	0.020187
물 마시는 빈도	0.010532
불면증 여부	0.009868
음주 여부	0.009983
긴장을 풀기 위해 다른 진정제나 약을 먹는다	0.015115
공복 글루코스	0.01011
당화혈색소	0.006686

유의미한 변수 : 10개

여자

변수	p-value
나이	0.004118
흡연여부	0.006381
당뇨병 진단 여부	0.018608
직업 종류	0.002839
불면증 여부	0.008693
긴장을 풀기 위해 진정제나 다른 약을 먹는다	0.013126
공복 글루코스	0.011057
당화혈색소	0.00634
BMI	0.005503
근육량	0.005379
체지방률	0.007384
총 콜레스테롤	0.004536

유의미한 변수 : 12개

앞서 유의미한 P-VALUE으로 추출 된 변수 대상 2차 통계 분석 실시

Part 4 >> 머신러닝 모델 분석

통계분석 - 남자

변수	p-value
음주기간	0.00399
고혈압 진단 여부	0.011257
당뇨병 진단 여부	0.017507
고지혈증 진단 여부	0.020187
물 마시는 빈도	0.010532
불면증 여부	0.009868
음주 여부	0.009983
긴장을 풀기 위해 다른 진정제나 약을 먹는다	0.015115
공복 글루코스	0.01011
당화혈색소	0.006686

통계분석 - 여자

변수	p-value
나이	0.004118
흡연여부	0.006381
당뇨병 진단 여부	0.018608
직업 종류	0.002839
불면증 여부	0.008693
긴장을 풀기 위해 진정제나 다른 약을 먹는다	0.013126
공복 글루코스	0.011057
당화혈색소	0.00634
BMI	0.005503
근육량	0.005379
체지방률	0.007384
총 콜레스테롤	0.004536

통계분석 P-value 유의한 값 -> 남자 10개 , 여자 12개

통계적 분석 VS 머신러닝 분석

비교를 위해 변수의 개수를 동일하게

남자 10개, 여자 12개로 통일

Part 4 >> 머신러닝 모델 분석 - 랜덤포레스트

남자(정확도: 0.9965)

변수	중요도
당뇨병 진단받은 경험유무	0.07782
비만 진단 - 체지방율	0.013356
평소 수면시간	0.012168
청량음료(콜라,사이다) 섭취빈도	0.011765
사탕/쵸콜렛 섭취빈도	0.0103
라면 섭취빈도	0.010275
직업 종류	0.009444
연령	0.009154
단 음식의 선호도	0.009004
비만 진단 - BMI	0.00885

여자(정확도: 0.9949)

변수	중요도
당뇨병 진단받은 경험유무	0.07107
중성지방	0.012348
평소 수면시간	0.012094
중성지방	0.011752
사탕/쵸콜렛 섭취빈도	0.011373
라면 섭취빈도	0.009211
단 음식의 선호도	0.009188
청량음료(콜라,사이다) 섭취빈도	0.007915
짠 음식의 선호도	0.007912
비만 진단 - BMI	0.007318
연령	0.007162
직업 종류	0.006863

Part 4 >> 머신러닝 모델 분석 - XGBOOST

남자(정확도: 0.99)

변수	중요도
공복 인슐린	0.029182
스트레스(여가시간을 통한 긴장 해소)	0.012709
스트레스(긴장해소를 위한 진정제 복용여부)	0.010305
스트레스(긴장 시 흡연여부)	0.008529
최근 피로를 자주 느끼는지 여부	0.007966
평소 수면시간	0.007545
현재 흡연 여부	0.00706
결혼상태	0.00676
월 평균 수입	0.006337
스트레스(신경성 소화불량)	0.00625

여자(정확도: 1.00)

변수	중요도
스트레스(긴장해소를 위한 진정제 복용여부)	0.043938
체지방율	0.017507
지난 한달간 체중 변화 여부	0.015714
피자/햄버거 섭취빈도	0.014997
현재 흡연 여부	0.014882
체지방 측정치	0.013302
하루 일과 중 육체적 활동시간 (경동활동)	0.010869
연령	0.008694
사탕/쵸콜렛 섭취빈도	0.007923
BMI	0.007785
스트레스(불면증 여부)	0.007488
근육량	0.00536

Part 4 >> 머신러닝 모델 분석 - CATBOOST

남자(정확도: 1.00)

변수	중요도
HDL콜레스테롤	4.900232
총 콜레스테롤	4.08095
지난 한달간 체중 변화 여부	4.069418
라면 섭취빈도	4.013711
인슐린(공복)	3.304795
현재 흡연 여부	1.808323
청량음료(콜라,사이다) 섭취빈도	1.805847
누적 흡연 담배량 (Pack-years)	1.180102
결혼상태	1.156945
하루 일과 중 육체적 활동시간 (중등활동)	1.051625

여자(정확도: 0.99)

변수	중요도
라면 섭취빈도	2.047092
치아상태	1.927034
하루 일과 중 육체적 활동시간 (중등활동)	1.809652
사탕/쵸콜렛 섭취빈도	1.589868
교육수준	1.491423
총 콜레스테롤	1.256798
체성분 분석 - 근육량	1.200099
트리글리세라이드(mg/dL)	1.122712
인슐린(공복)	1.110858
음주여부	1.069346
스트레스(신경성 소화불량)	1.046795
음주 기간	1.016632

Part 4 >> 머신러닝 모델 분석 - LightGBM

남자(정확도: 0.9928)

변수	중요도
청량음료(콜라,사이다) 섭취빈도	146
라면 섭취빈도	127
단 음식 선호도	121
월 평균 수입	95
교육수준	76
사탕/쵸콜렛 섭취빈도	74
치아 상태	69
피자/햄버거 섭취빈도	68
짠 음식 선호도	66
직업 종류	64

여자(정확도: 0.9966)

변수	중요도
사탕/쵸콜렛 섭취빈도	88
중동 활동	87
수면 시간	83
라면 섭취빈도	78
청량음료(콜라,사이다) 섭취빈도	75
치아 상태	67
월 평균 수입	65
연령	64
단 음식 선호도	60
총 콜레스테롤	58
최근 피로를 자주 느끼는지 여부	55
체지방 측정치	54

Part 4 >> 머신러닝 모델 분석 - GradientBoosting

남자(정확도: 0.9946)

변수	중요도
청량음료 섭취빈도	0.0034
공복 인슐린	0.0027
평균 수입	0.0026
결혼 상태	0.002
긴장을 풀기 위해 진정제나 다른 약을 먹는다	0.0017
피자/햄버거 섭취빈도	0.0009
당뇨병 진단 여부	0.0008
밤,주말과 같은 여가시간을 통해 긴장을 푼다	0.0008
라면 섭취빈도	0.0007
단 음식 선호도	0.0006

여자(정확도: 0.9949)

변수	중요도
피자/햄버거 섭취빈도	0.0066
체지방률	0.0028
긴장을 풀기 위해 진정제나 다른 약을 먹는다	0.0025
체지방 측정치	0.0024
짠 음식 선호도	0.0013
나이	0.0012
총 콜레스테롤	0.001
사탕/쵸콜렛 섭취빈도	0.001
공복 인슐린	0.0008
결혼 상태	0.0008
짠 음식 선호도	0.0006
BMI	0.0005

Part 4 >> 머신러닝 모델 분석 - Adaboost

남자(정확도: 0.9946)

변수	중요도
청량음료(콜라,사이다) 섭취빈도	0.1
평소 수면시간	0.08
단 음식의 선호도	0.08
월 평균 수입	0.06
긴장감, 두통,목과 어깨의 통증,불면증이 있다	0.06
하루 일과 중 육체적 활동시간 (격한활동)	0.04
긴장했을때 먹거나 마시거나 담배를 피운다	0.04
불면증 여부	0.04
음주 기간	0.04
사탕/쵸콜렛 섭취빈도	0.04

여자(정확도: 0.9949)

변수	중요도
평소 수면시간	0.08
치아상태	0.06
라면 섭취빈도	0.06
단 음식의 선호도	0.06
피자/햄버거 섭취빈도	0.06
사탕/쵸콜렛 섭취빈도	0.04
스트레스(긴장감, 두통,목과 어깨의 통증,불면증이 있다)	0.04
체지방율	0.04
직업 종류	0.04
밤,주말과 같은 여가시간을 통해 긴장을 푼다	0.04
하루 일과 중 육체적 활동시간 (중동활동)	0.04
원래 술을 못 마시거나 또는 처음부터 술을 안 마십니까	0.02

Part 4 >> 머신러닝 모델 분석 – MFI(Mean Feature Importance)

MFI 남자

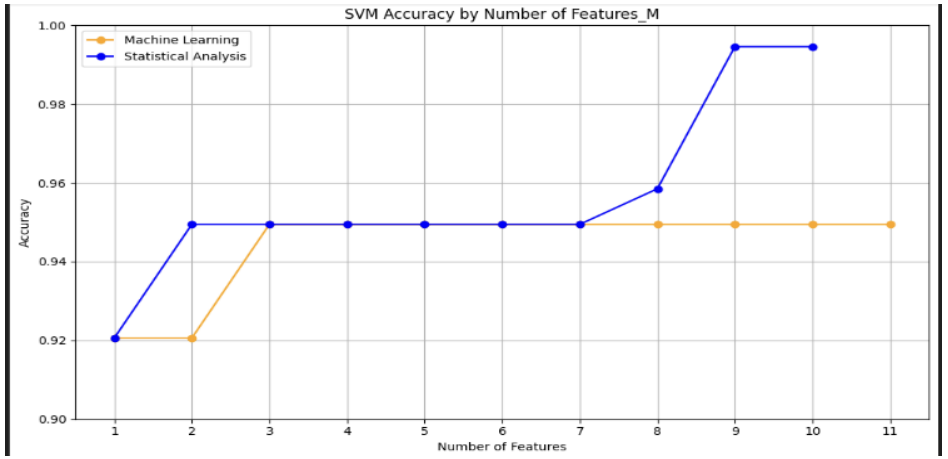
변수	중요도
라면 섭취빈도	62.310747
체지방률	57.519355
당뇨병 진단 여부	55.576267
평소 수면시간	53.734483
BMI	53.117428
단 음식 선호도	50.432465
사탕/초콜렛 섭취빈도	43.54519
직업 종류	41.179109
교육 수준	37.722343
청량음료 섭취빈도	36.016645

MFI 여자

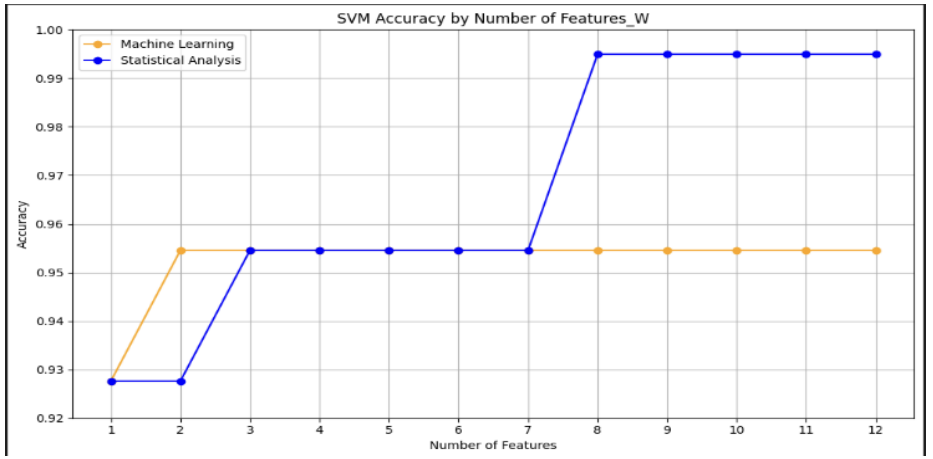
변수	중요도
직업 종류	64.933567
당뇨병 진단 여부	55.585305
총 콜레스테롤	51.583617
사탕/초콜렛 섭취빈도	50.055309
평소 수면시간	46.892015
짠 음식 선호도	43.818315
청량음료 섭취빈도	41.780891
중성지방	39.864399
BMI	36.32913
라면 섭취빈도	32.608435
나이	29.654554
단 음식 선호도	24.227797

Part 5 >> SVM & 로지스틱

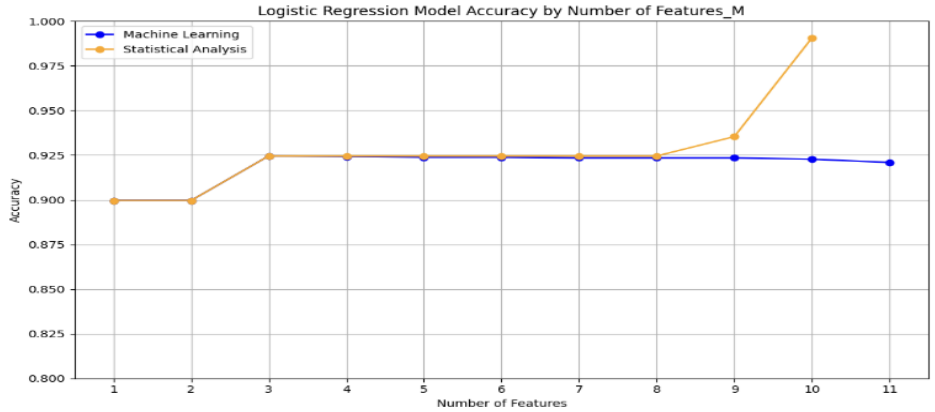
SVM 남자



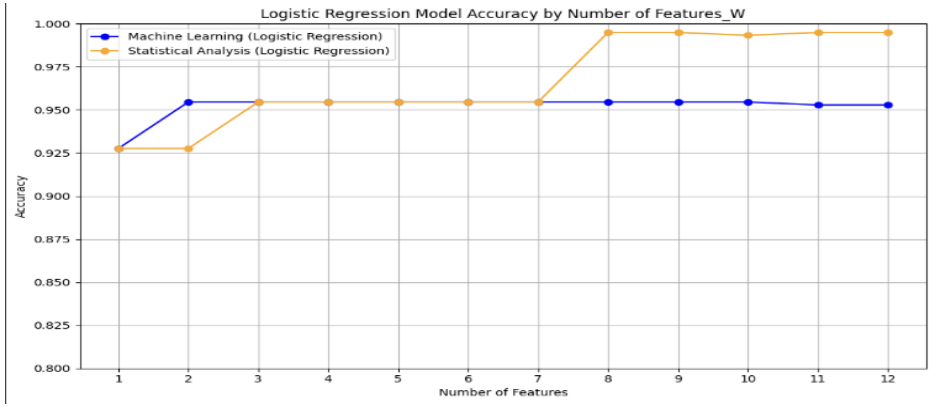
SVM 여자



로지스틱 남자



로지스틱 여자



데이터 차원 축소 후 SVM & 로지스틱의 정확도 측정

➔ 남자, 여자 모두 통계적 분석 시 당뇨병 예측에 효과적

Part 5 >> 최종 결론

➔ 남자 최종 추출 변수

1. 음주기간
2. 고혈압 진단 여부
3. 당뇨병 진단 여부
4. 고지혈증 진단 여부
5. 물 마시는 빈도
6. 불면증 여부
7. 음주 여부
8. 긴장을 풀기 위해 다른 진정제나 약을 먹는다
9. 공복 글루코스
10. 당화혈 색소

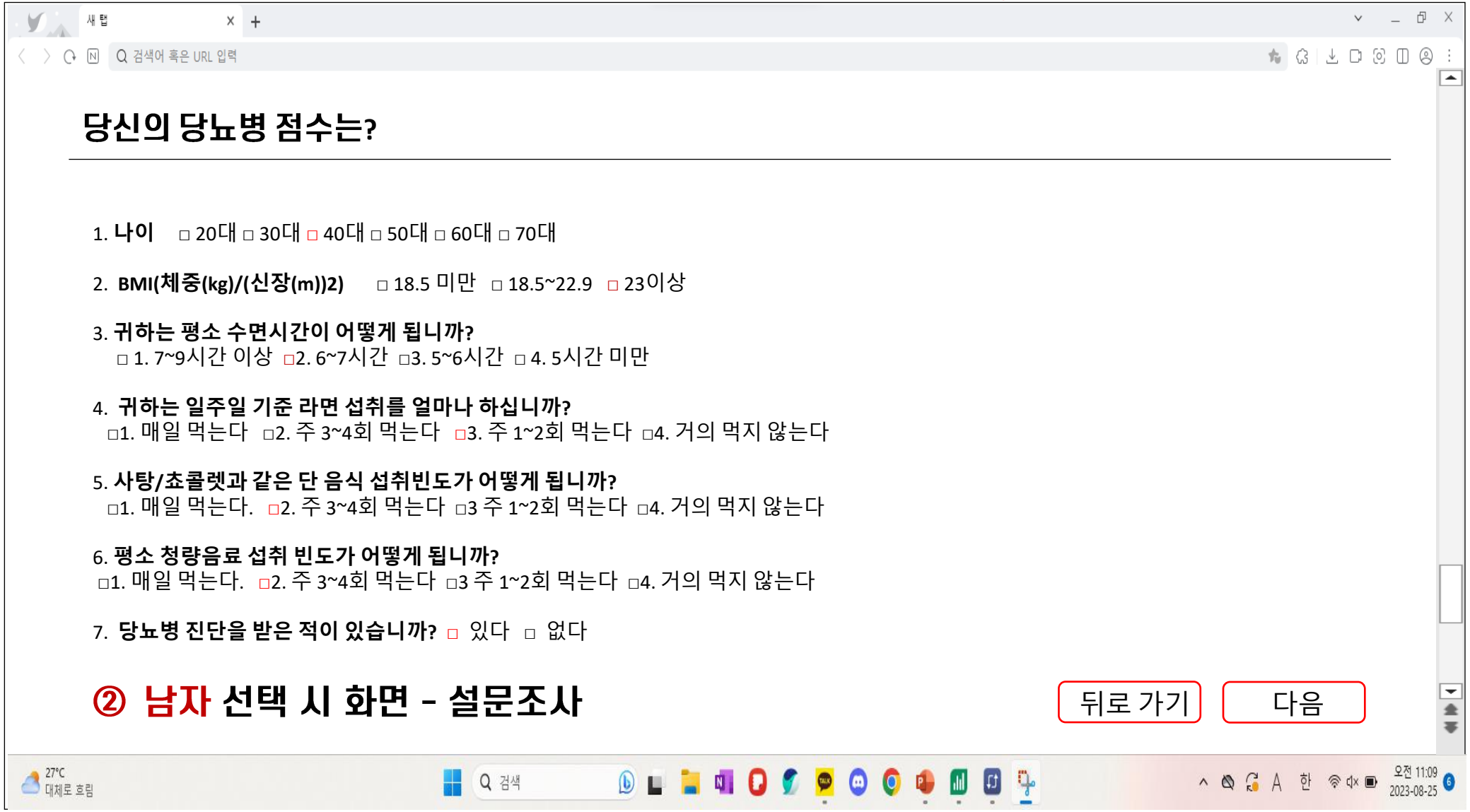
➔ 여자 최종 추출 변수

1. 나이
2. 흡연 여부
3. 당뇨병 진단 여부
4. 직업 종류
5. 불면증 여부
6. 긴장을 풀기 위해 진정제나 다른 약을 먹는다
7. 공복 글루코스
8. 당화혈 색소
9. BMI
10. 근육량
11. 체지방률
12. 총 콜레스테롤

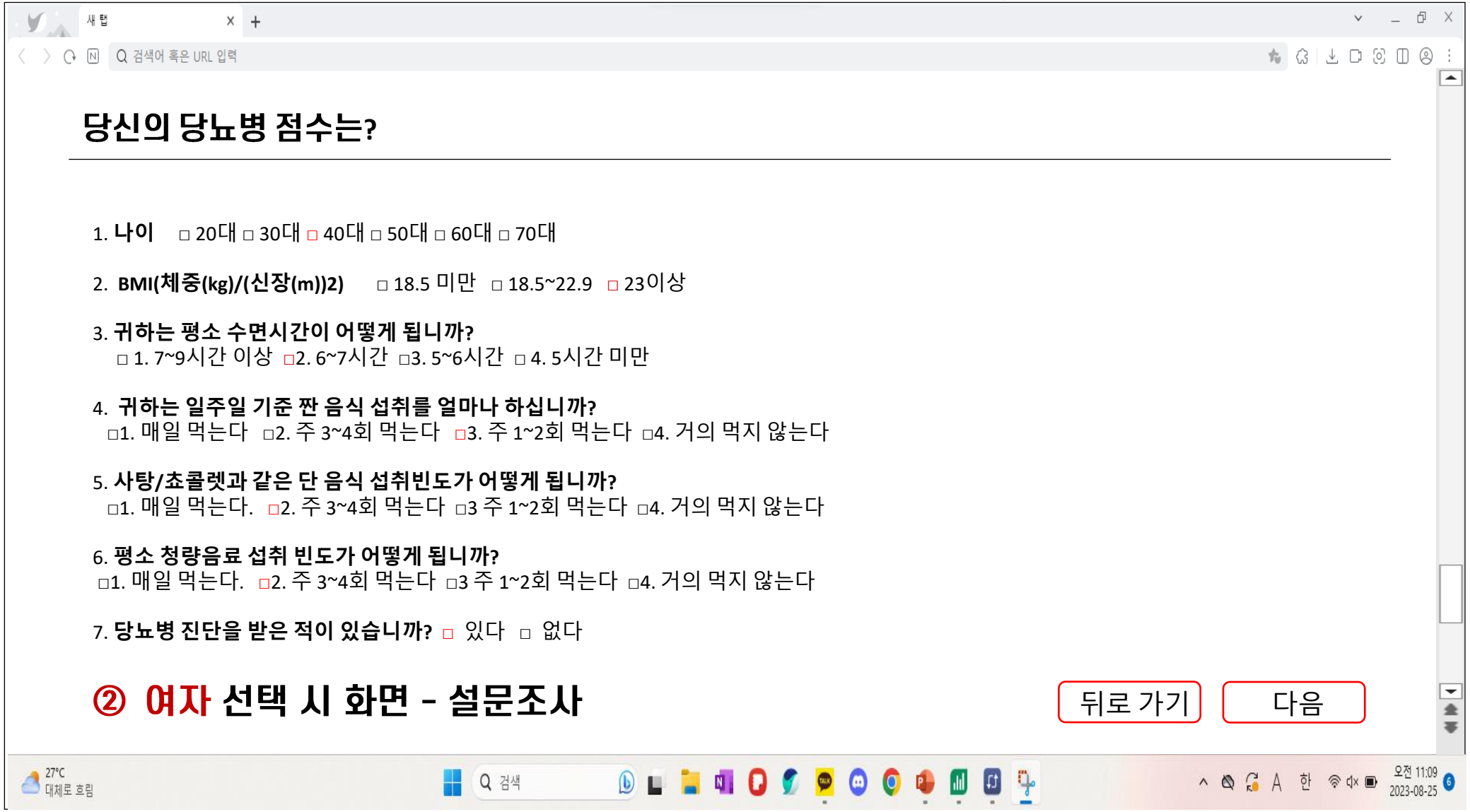
Part 5 >> 최종 결론 – 앱 구현

①성별 선택화면

Part 5 >> 최종 결론 – 앱 구현



Part 5 >> 최종 결론 – 앱 구현



Part 5 >> 최종 결론 – 앱 구현