# 서울시 먹거리 통계조사를 활용한 건강 상태 분석

7기 데이터 분석 중급반
김인서37

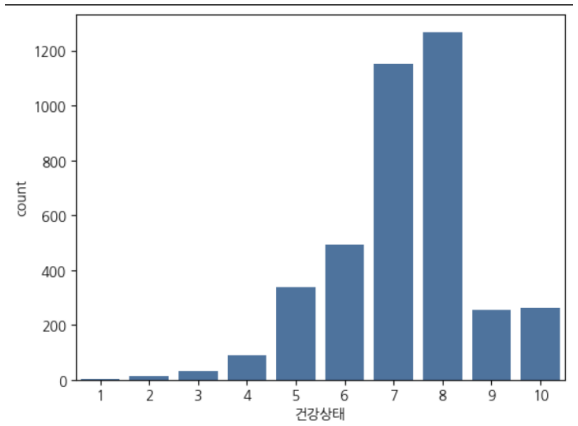
# 활용 데이터와 사용할 컬럼

- A_SQ4C1: 성별(1: 남자, 2: 여자)
- DE2: 연령대(만 18~29, 30대, 40대, 50대, 60대, 70대 이상)
- B6_1, B6_2: 키, 몸무게
- Q1_1 ~ Q1_4: 아침, 점심, 저녁, 합계 식사 횟수(최근 1주)
- Q4: 채식 육식 여부, 숫자가 클수록 육식 성향, 작을수록 채식 성향
- Q6_1 ~ Q6_12: 전곡류, …, 유제품 식사 횟수 (최근 1년), 숫자가 클수록 횟수가 많음
- Q7_1 ~ Q7_4: 가당음료, …, 주류 식사 횟수(최근 1년), 숫자가 클수록 횟수가 많음
- Q15: 건강 상태 점수(0~10), target

서울시 먹거리 통계조사 2022

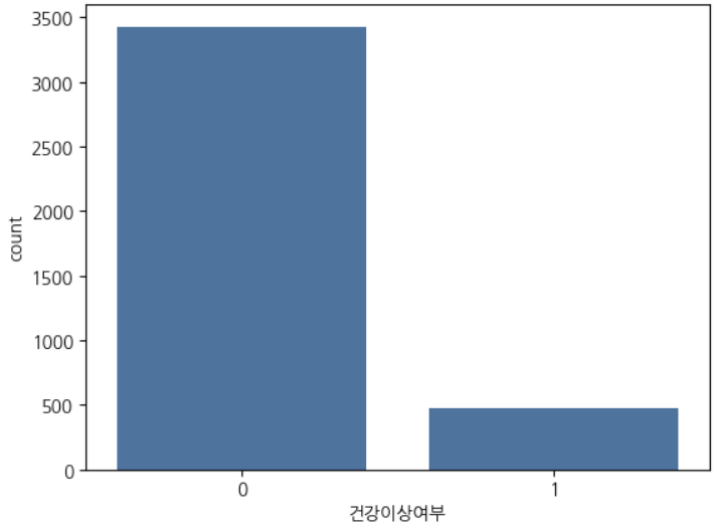
서울시 먹거리 통계조사> 데이터셋> 공공데이터 | 서울열린데이터광장 (seoul.go.kr)

# 데이터 확인




Target 분포가 다음과 같으므로 1~5점을 1, 6~10점을 0으로 설정
(1을 탐지하는 것이 목표)

이를 '건강이상여부'라는 새로운 컬럼으로 설정하면
1의 비율이 15% 미만인 불균형이 심한 데이터가 된다.



# 데이터 전처리

- 영문 코드로 되어 있는 컬럼을 설명에 맞게 한글로 변환
- 키와 몸무게를 이용해 BMI, 저체중, 비만 여부 계산 후 키, 몸무게 컬럼 삭제
- 성별을 0: 남자, 1: 여자로 변환

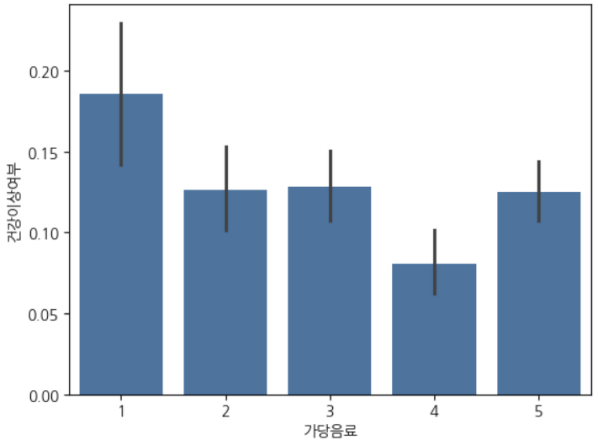
➔ 다음과 같은 3904개의 행과 27개의 컬럼이 나온다.

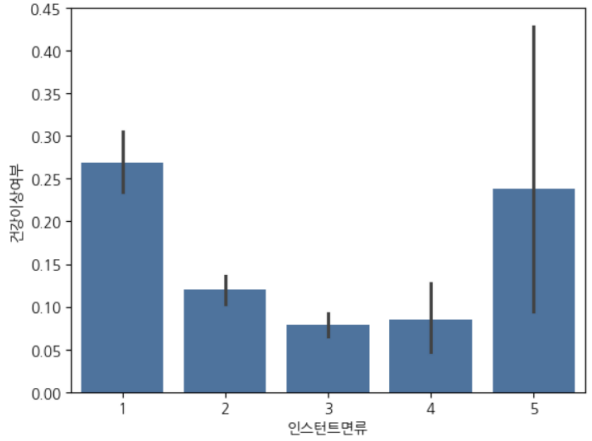
| | 성별 | 연령대 | 아침 | 점심 | 저녁 | 합계 | 채식육식여부 | 전곡류 | 생채소류 | 채소반찬류 | ... | 생과일류 | 유제품 | 가당음료 | 인스턴트면류 | 패스트푸드 | 주류 | 건강이상여부 | BMI | 저체중 | 비만 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 5 | 7 | 7 | 7 | 21 | 5 | 1 | 9 | 9 | ... | 8 | 3 | 2 | 3 | 1 | 3 | 0 | 21.513859 | 0 | 0 |
| **1** | 1 | 5 | 7 | 7 | 7 | 21 | 5 | 8 | 8 | 9 | ... | 4 | 3 | 2 | 2 | 1 | 2 | 0 | 24.034610 | 0 | 0 |
| **2** | 0 | 2 | 5 | 7 | 7 | 19 | 1 | 3 | 5 | 5 | ... | 4 | 5 | 2 | 3 | 3 | 4 | 0 | 23.120624 | 0 | 0 |
| **3** | 1 | 2 | 5 | 7 | 7 | 19 | 1 | 4 | 5 | 3 | ... | 3 | 5 | 2 | 3 | 2 | 2 | 0 | 21.484375 | 0 | 0 |
| **4** | 0 | 3 | 0 | 7 | 7 | 14 | 6 | 3 | 1 | 2 | ... | 4 | 5 | 5 | 4 | 3 | 3 | 1 | 27.471689 | 0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **3899** | 1 | 3 | 5 | 6 | 7 | 18 | 5 | 7 | 7 | 8 | ... | 5 | 6 | 5 | 3 | 2 | 2 | 0 | 21.936347 | 0 | 0 |
| **3900** | 1 | 5 | 3 | 7 | 5 | 15 | 5 | 5 | 8 | 7 | ... | 7 | 4 | 3 | 3 | 2 | 1 | 0 | 22.481329 | 0 | 0 |
| **3901** | 0 | 5 | 5 | 6 | 5 | 16 | 5 | 8 | 7 | 6 | ... | 8 | 4 | 4 | 2 | 1 | 2 | 0 | 24.622961 | 0 | 0 |
| **3902** | 1 | 2 | 5 | 6 | 3 | 14 | 5 | 7 | 8 | 7 | ... | 8 | 7 | 4 | 3 | 3 | 2 | 0 | 20.202020 | 0 | 0 |
| **3903** | 1 | 4 | 6 | 7 | 7 | 20 | 5 | 8 | 8 | 9 | ... | 5 | 3 | 2 | 2 | 2 | 2 | 1 | 24.524346 | 0 | 0 |

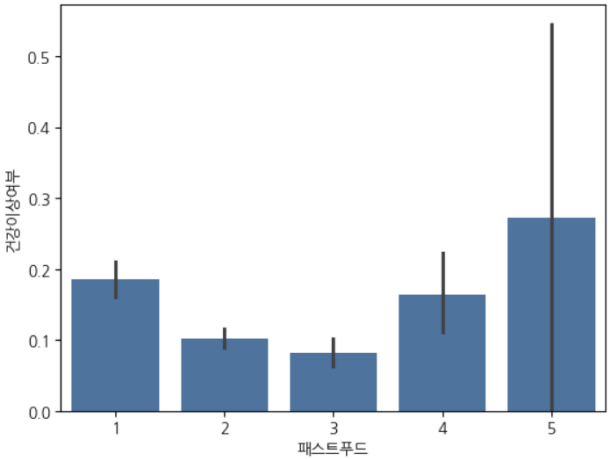
3904 rows × 27 columns

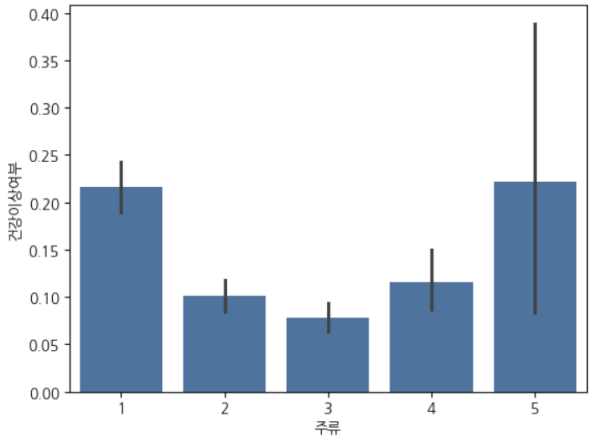
# 가설 설정, 검정

가설 1: 가당음료, 인스턴트면류, 패스트푸드, 주류 소비량이 많을수록 건강 이상 응답률이 높다.



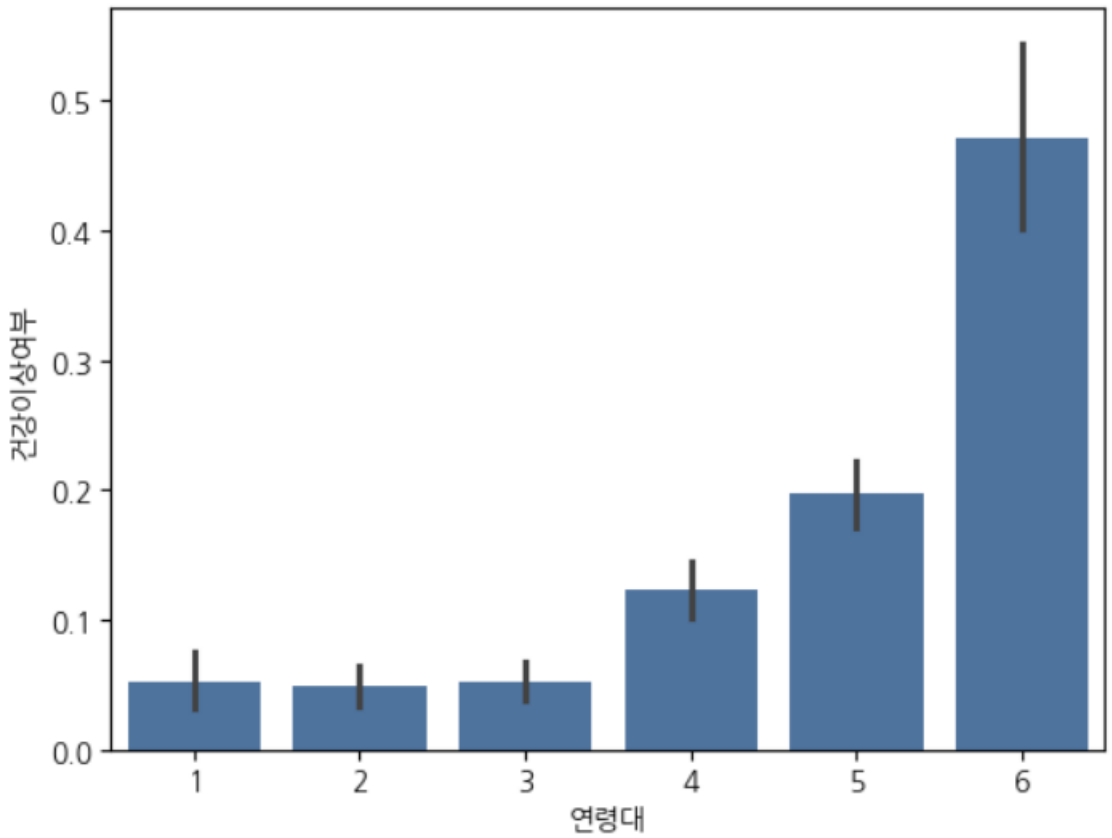

결론:

소비량이 많다고 건강 이상 응답률이 높은 것은 아니다.

너무 적게 소비하거나 너무 많이 소비하는 것보다는 적당한 빈도로 소비할 때 가장 건강 이상 응답률이 낮은 것으로 보인다.

# 가설 설정, 검정

가설 2: 연령대가 높을수록 건강 이상 응답률이 높다.



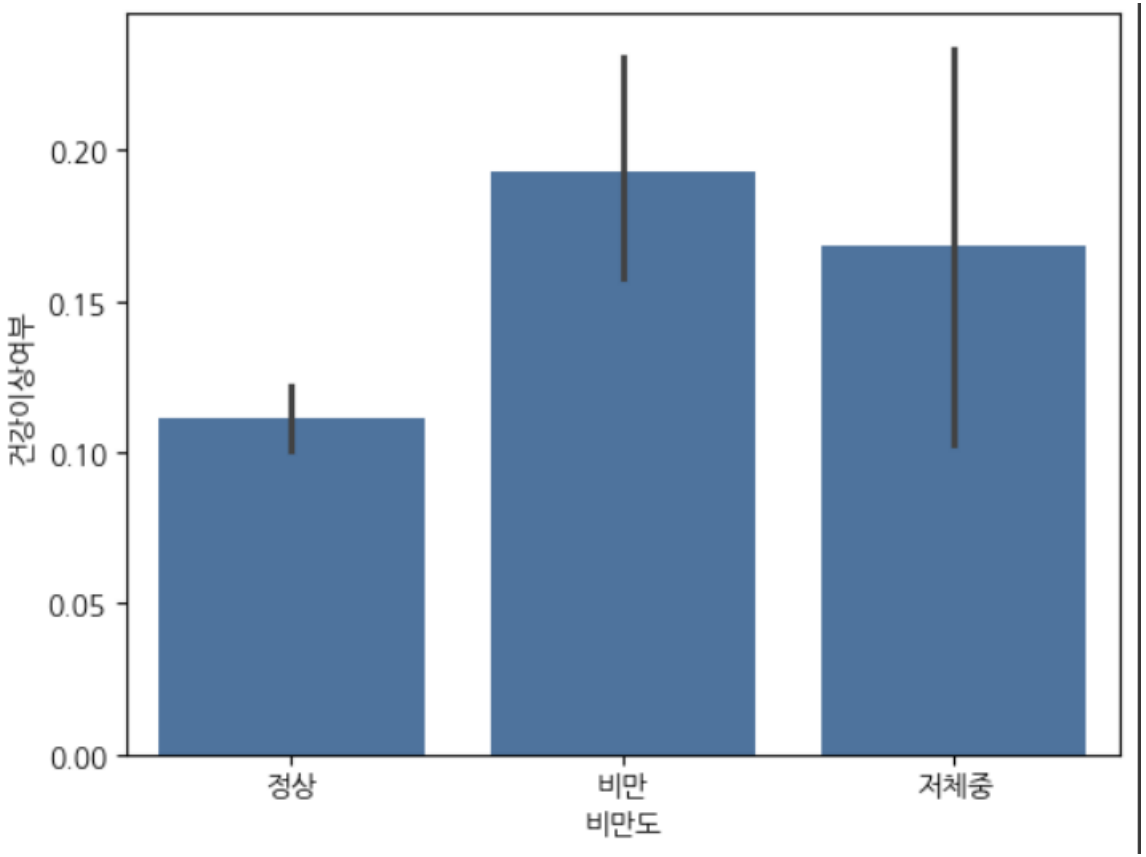

결론:

연령대가 높을수록 건강 이상 응답률이 급격하게 증가했으며,

특히 6(70대 이상)집단에서는 건강 이상 응답률이 50% 정도에 달했다.

# 가설 설정, 검정

가설 3: BMI가 18.5 미만인 저체중, 25 이상인 비만은 정상 체중에 비해 건강 이상 응답률이 높을 것이다.



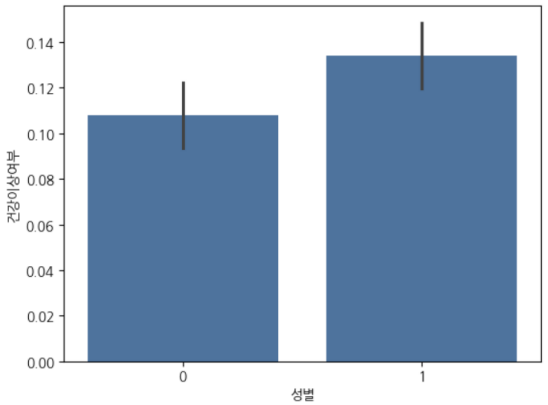

결론:

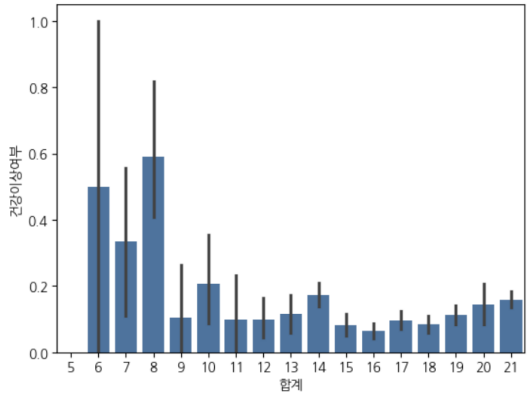
신뢰구간을 고려했을 때, 저체중과 비만 간에는 명확한 차이가 드러나지 않지만,

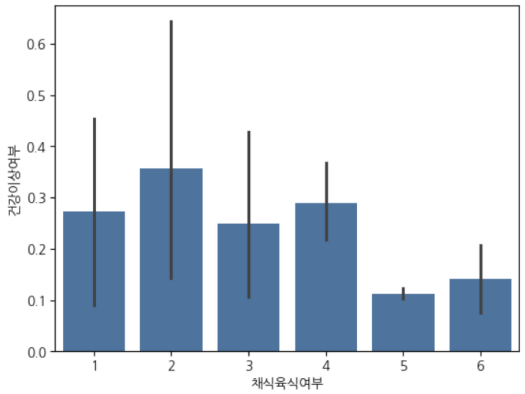
정상과 비만, 정상과 저체중 간에는 차이가 드러남을 알 수 있다.
(두 경우 모두 정상 체중인 사람이 건강 이상 응답률이 낮다.)

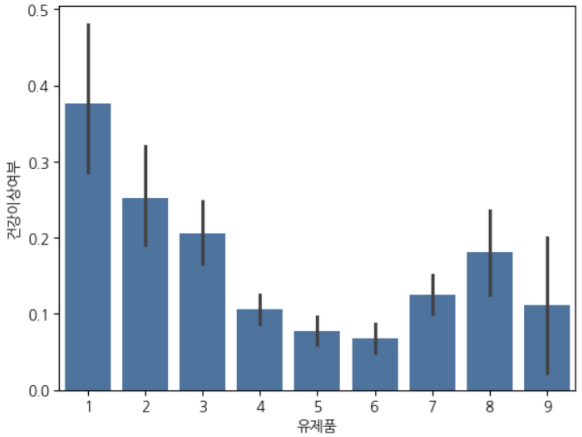
# 시각화

남자보다는 여자가, 특정 횟수 이하로 식사하는 경우, 육식보다는 채식이 건강 이상 응답 비율이 상대적으로 높았다.




음식 섭취 빈도 설문에서는 전반적으로 너무 적은 빈도로 섭취하거나 너무 많은 빈도로 섭취하지 않고 적절한 빈도로 섭취했을 때 건강 이상 응답 비율이 가장 적은 것으로 파악되었다.

# 모델링

X는 target에 해당하는 '건강이상여부'를 제외한 나머지 26개의 컬럼
Y는 target에 해당하는 '건강이상여부'로 분리

train, test는 y의 비율을 보존하면서 각각 70%, 30%로 분리

```
1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
```

# 모델링

Target의 불균형이 심하므로 균형을 맞춰야 하는데,
언더 샘플링을 할 경우 데이터 손실량이 너무 많으므로
오버 샘플링 중 SMOTE 방법 활용

SMOTE를 활용하면 Random Oversampling보다 과적합을 줄일 수 있음.

```
[  1 from imblearn.over_sampling import SMOTE
   2
   3 sm = SMOTE(random_state=42)
   4 X_train_sm, y_train_sm = sm.fit_resample(X_train, y_train)
```

```
[  1 from collections import Counter
   2
   3 print(Counter(y_train_sm))
```

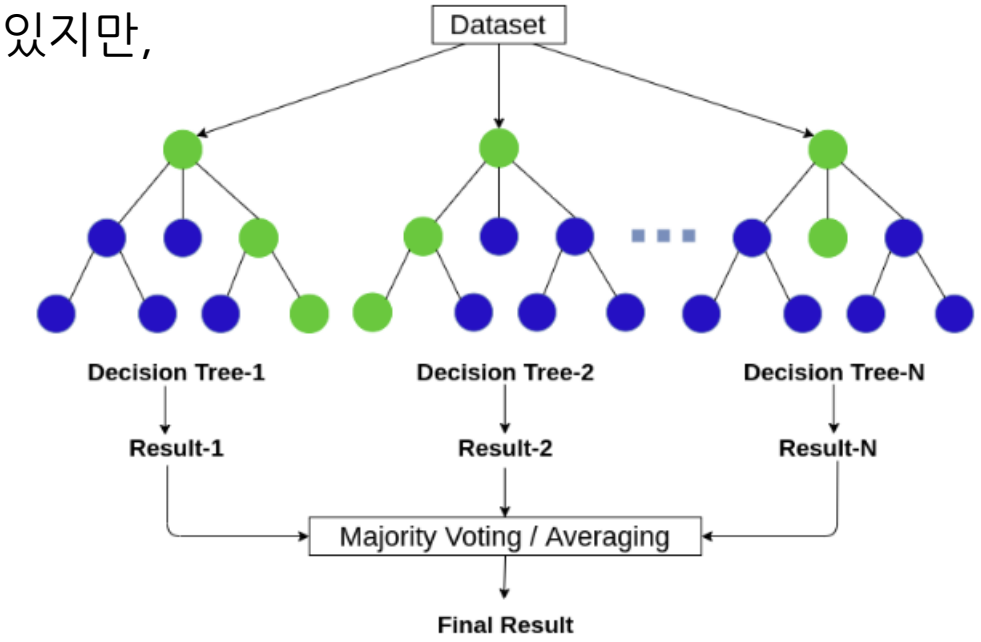
```
Counter({0: 2399, 1: 2399})
```

# 모델링

사용 모델: RandomForestClassifier

Decision Tree에 Bagging(Bootstrap aggregation) 앙상블 기법을 활용한 모델

일반적으로 과적합을 해소하기 유리해 정확도가 높다는 장점이 있지만,
계산 비용이 높고 추론 로직을 설명하기 어려운 단점이 있음.



# 모델링

## 모델링 결과

RandomForestClassifier(max_depth=12)일 때
F1_macro가 0.70이 나옴을 확인할 수 있다.
(튜닝은 수작업으로 진행)

```
1 rf_clf_final = RandomForestClassifier(max_depth=12, random_state=42)
2 rf_clf_final.fit(X_train_sm, y_train_sm)
```

RandomForestClassifier
RandomForestClassifier(max_depth=12, random_state=42)

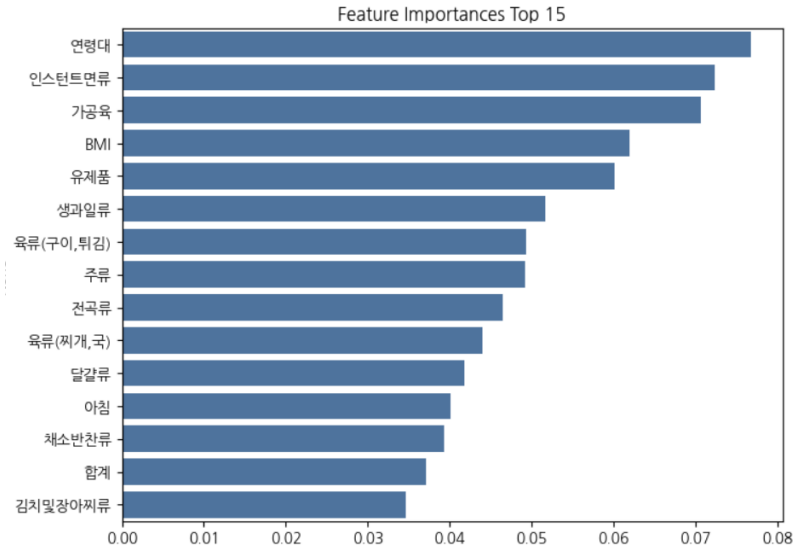
```
1 pred_final = rf_clf_final.predict(X_test)
2
3 print(classification_report(y_test, pred_final))
4 print(confusion_matrix(y_test, pred_final))
```

```
              precision    recall  f1-score   support

           0       0.93      0.93      0.93      1029
           1       0.47      0.47      0.47       143

    accuracy                           0.87      1172
   macro avg       0.70      0.70      0.70      1172
weighted avg       0.87      0.87      0.87      1172

[[953  76]
 [ 76  67]]
```




Feature Importance를 확인해 봤을 때 '연령대'가 모델 분류에 가장 많은 관여를 했음을 알 수 있다.

# 결론과 한계점

**결론**

모델링을 통해
- 건강 이상 여부에는 연령대가 가장 중요한 영향을 줌을 확인

시각화를 통해
- 식습관(음식 종류 섭취 빈도)에서는 너무 많은 빈도나 적은 빈도가 아닐 때 가장 건강 이상 빈도가 낮음을 확인
- 식사 횟수가 너무 적거나, 채식 위주, 저체중 또는 비만인 경우에는 건강 이상 위험이 더 높음

- 이를 바탕으로 고연령자를 중점적으로 채식주의자, 식사 횟수가 너무 적은 사람, 저체중 또는 비만인 사람에 대해 모니터링을 하면 국민 건강 증진에 도움을 줄 수 있을 것으로 판단됨.

**한계점**

- 개인의 건강 응답에 주관이 들어가 있으므로 응답에 편향이 있을 수 있음
- 모델링 결과 0에 대한 F1 score는 0.93, 1에 대한 F1 score는 0.47로 1을 상대적으로 잘 감지하지 못했음

감사합니다.