서울시 먹거리 통계조사를 활용한 건강 상태 분석

7기 데이터 분석 중급반
김인서 37

I. 서론

1. 수집 데이터

서울시 먹거리 통계조사 2022 (서울 열린데이터광장)

2. 분석 목적

서울시 먹거리 통계조사 데이터를 활용해 건강 이상에 영향을 주는 식습관이나 식습관 외의 요소가 무엇인지 파악하고 분석하고자 하였다.
또한, 여러 컬럼들을 활용해 건강 이상을 파악할 수 있는 모델을 만들고자 하였다.

3. 사용한 컬럼

여러 컬럼 중 건강과 관련이 있는 요소만 추출하여서 분석에 활용하였다.

A_SQ4C1
- 성별 (1: 남자, 2: 여자)

DE2
- 연령대 (1: 만 18~29세, 2: 30대, 3: 40대, 4: 50대, 5: 60대, 6: 70대 이상)

B6_1, B6_2
- 각각 키(cm), 몸무게(kg)

Q1_1 ~ Q1_4
- 각각 최근 1주일간 아침, 점심, 저녁, 도합 식사 횟수

Q4
채식, 육식 여부
(1 ~ 6까지 있으며, 1로 갈수록 채식의 정도가, 6으로 갈수록 육식의 정도가 강하다.)

Q6_1 ~ Q6_12
최근 1년간 전곡류, ... , 유제품 식사 횟수
(1 ~ 9까지 있으며, 숫자가 커질수록 자주 섭취함을 의미한다.)

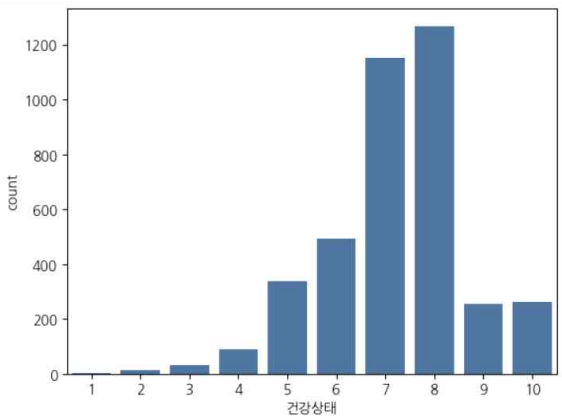
Q7_1 ~ Q7_4
최근 1년간 가당음료, ... , 주류 식사 횟수
(1 ~ 5까지 있으며, 숫자가 커질수록 자주 섭취함을 의미한다.)

Q15
건강 상태 점수
(가공해서 target으로 활용할 것이다.)

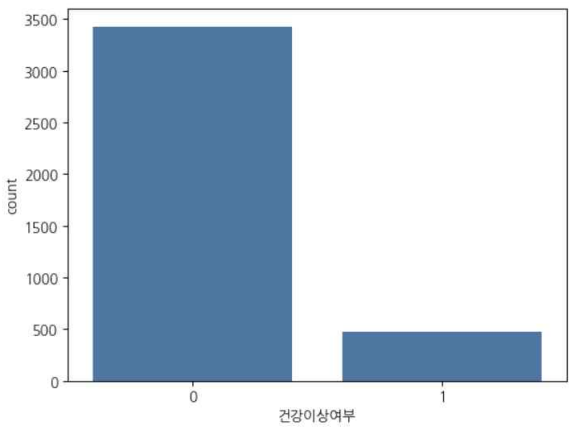
II. 본론

1. 데이터 확인

target으로 사용할 건강 상태 컬럼은 다음과 같은 분포를 따른다.



여기서 우리는 건강 이상 여부를 탐지하고 싶으므로 1 ~ 5를 1로, 6 ~ 10을 0으로 바꾼 후 이진 분류 문제로 풀 것이다.



이런 식으로 건강 이상 여부라는 새로운 컬럼으로 만들면 1의 비율이 15% 미만인 불균형이 심한 데이터가 만들어진다.

2. 데이터 전처리

컬럼 이름이 영어 코드로 표현이 되어 있기 때문에 눈에 보기 편하게 하기 위해서 컬럼 자체의 설명에 맞게 한글로 변환하는 과정을 거쳤다.

또한, 키와 몸무게를 이용해서 BMI를 계산하였고, 이를 이용해 저체중(BMI 18.5 미만), 비만(BMI 25 이상) 여부를 나타내는 컬럼을 생성하고, 원래의 키와 몸무게 컬럼은 제거하였다.

원래 성별에 있는 데이터는 남성이면 1, 여성이면 2인데, 이를 남성이면 0, 여성이면 1로 변환하였다.

이렇게 변환하면 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

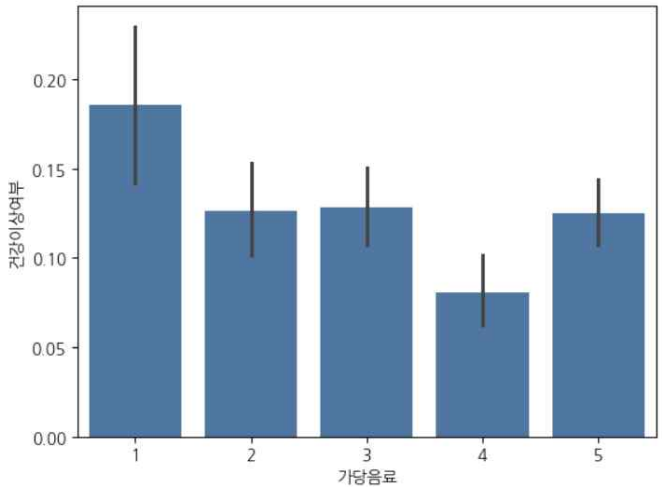
3904 rows × 27 columns

3. 가설 설정 및 검정

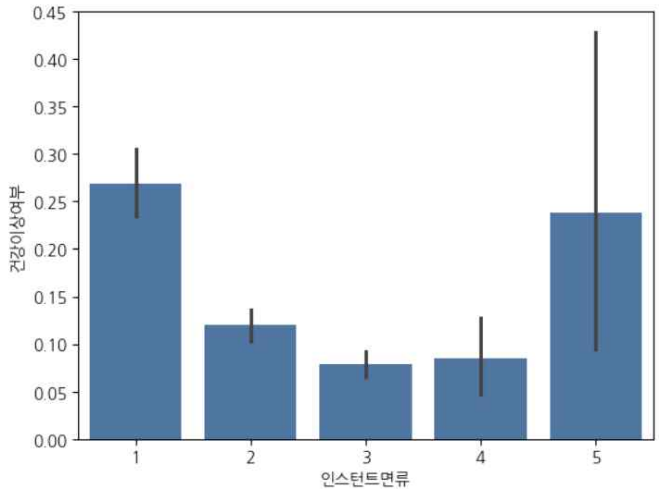
가설 1.

- 가당음료, 인스턴트 면류, 패스트푸드, 주류 소비량이 많을수록 건강 이상 응답률이 높을 것이다.

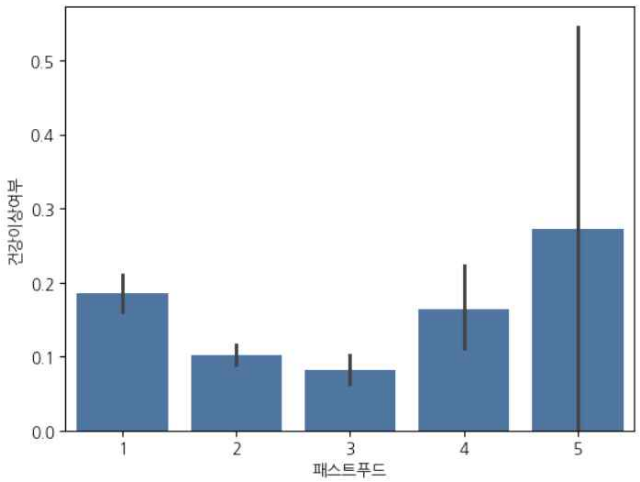
(가당음료)



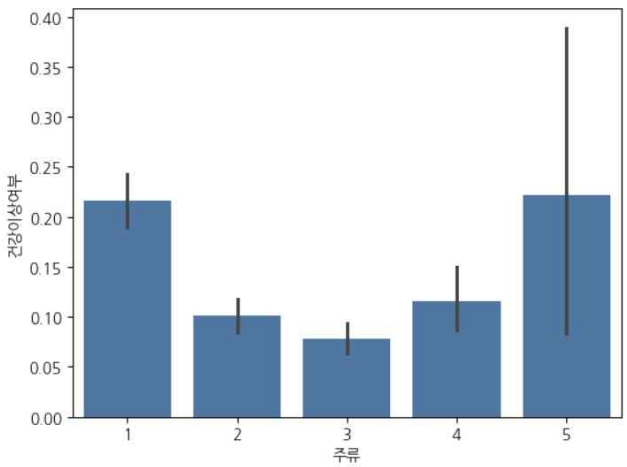
(인스턴트 면류)



(패스트푸드)



(주류)

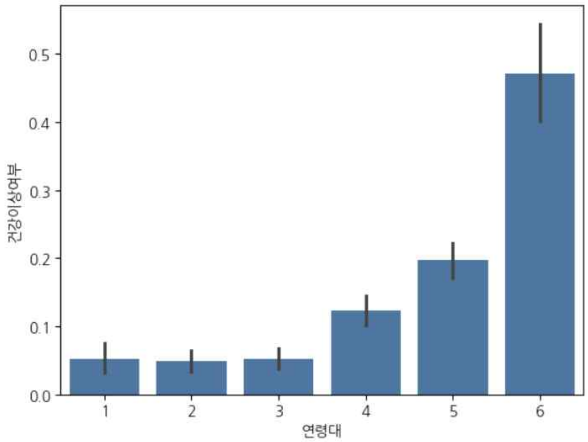


결론

- 전반적으로 가설대로의 경향이 드러나지 않았다.
- 적절한 빈도로 먹는 것이 그렇지 않은 것보다 건강 이상 빈도가 낮은 것으로 보인다.

가설 2.

- 연령대가 높을수록 건강 이상 응답률이 높을 것이다.

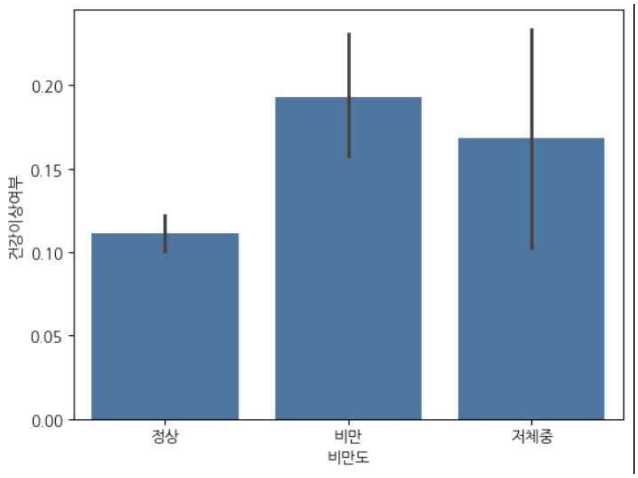


결론

- 실제로 연령대가 높을수록 건강 이상 응답률이 급격하게 증가했다.
- 특히 6집단(70대 이상)에서는 건강 이상 응답률이 50% 정도에 달했다.

가설 3.

- 저체중과 비만은 정상 체중에 비해 건강 이상 응답률이 높을 것이다.

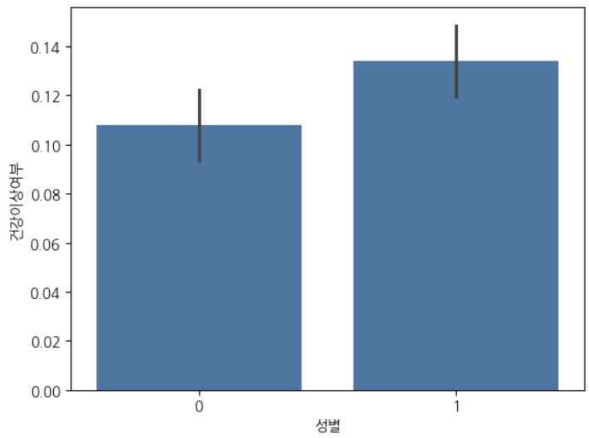


결론

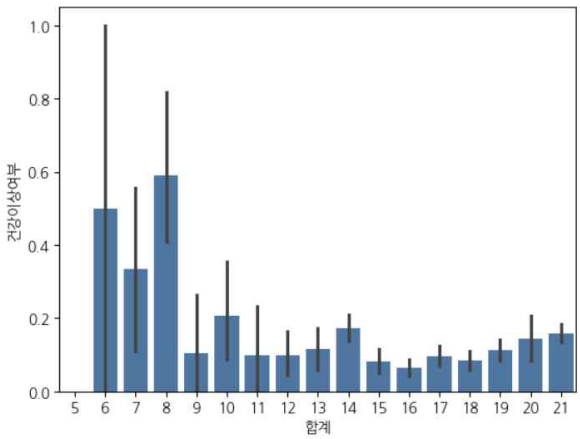
- 신뢰구간을 고려했을 때, 저체중과 비만 간에는 명확한 차이가 드러나지 않았지만, 정상과 비만, 정상과 저체중 간에는 차이가 드러남을 알 수 있다.

4. 기타 시각화

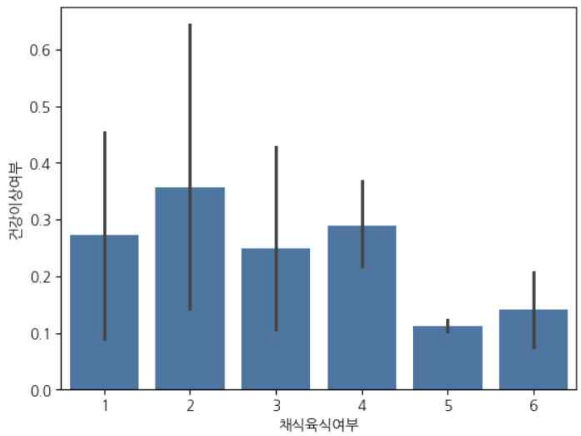
(성별)



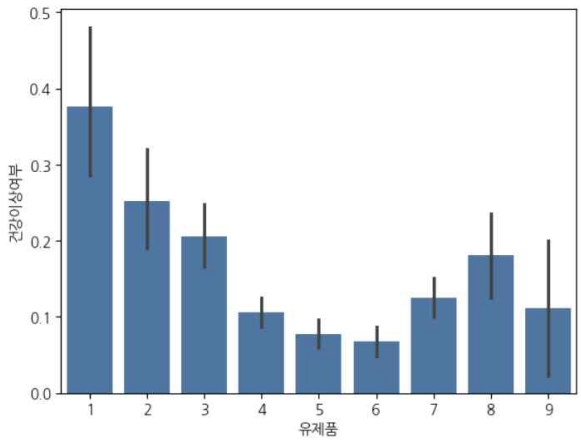
(식사 횟수 합계)



(채식, 육식 여부)



(음식 섭취 빈도 - 유제품)



결론

- 여자가 남자보다 건강 이상 응답률이 높았다.
- 1주일 동안 8회 이하 식사를 한 경우의 건강 이상 응답률이 두드러졌다.
- 채식주의자가 채식주의가 아닌 사람보다 건강 이상 응답률이 높았다.

- 음식 섭취 빈도 설문에서는 유제품뿐만 아니라 전반적으로 너무 적은 빈도로 섭취하거나 너무 많은 빈도로 섭취하지 않고 적절한 빈도로 섭취했을 때 건강 이상 응답 비율이 가장 적은 것으로 파악되었다.
- 그 이유는 한정된 식사 횟수에서 음식의 균형이 깨지게 먹으면 영양의 불균형이 생겨서 균형 있게 먹는 것보다 좋지 않기 때문인 것으로 추측된다.

5. 모델링

1) Train, Test 분리

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

X는 target인 건강 이상 여부를 제외한 나머지 26개의 컬럼
y는 target인 건강 이상 여부 컬럼으로 분리했다.

train, test는 y의 비율을 보존하면서 각각 70%, 30%로 분리했다.

2) 오버샘플링

```
from imblearn.over_sampling import SMOTE

sm = SMOTE(random_state=42)
X_train_sm, y_train_sm = sm.fit_resample(X_train, y_train)
```

```
from collections import Counter

print(Counter(y_train_sm))
```

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

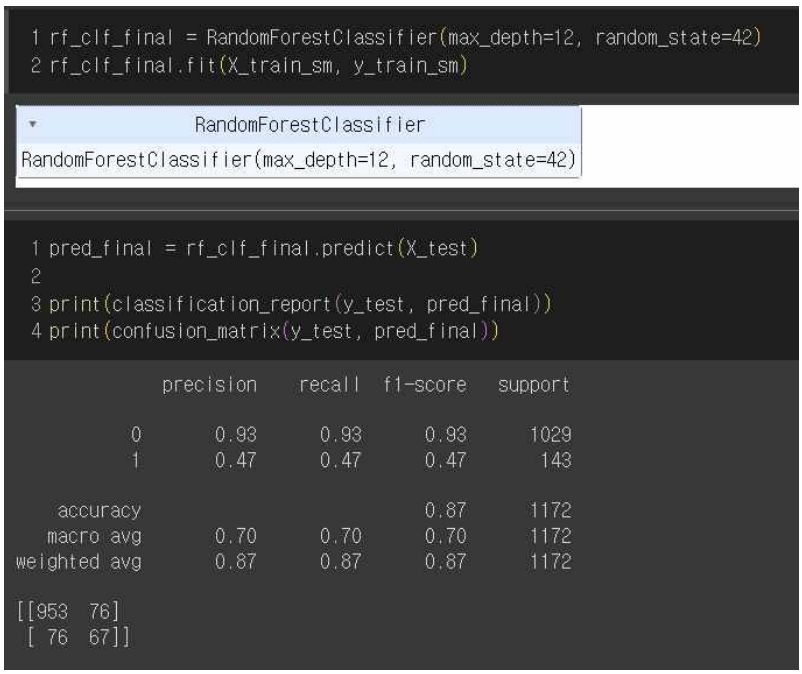
target의 불균형이 심하기 때문에 균형을 맞추어줄 필요가 있다.
언더샘플링을 하면 데이터가 너무 작아져서 정보 손실량이 크므로 오버샘플링 중 SMOTE 방법을 활용했다.

SMOTE를 활용하면 랜덤 오버샘플링보다 과적합을 줄일 수 있다.

3) 랜덤 포레스트, 피처 중요도 확인

또한, 여기서는 랜덤 포레스트 모델을 사용했는데, 랜덤 포레스트는 Decision Tree에 Bagging (Bootstrap aggregation) 앙상블 기법을 활용한 모델이다.
일반적으로 과적합을 해소하기 유리해서 정확도가 높다는 장점이 있지만, 계산 비용이 높고 추론 로직을 설명하기 어려운 단점이 있다.

여기서는 scikit-learn에 있는 RandomForestClassifier를 활용하였다.
튜닝은 정확도를 비교해가면서 수작업으로 진행하였다.



```
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))
```

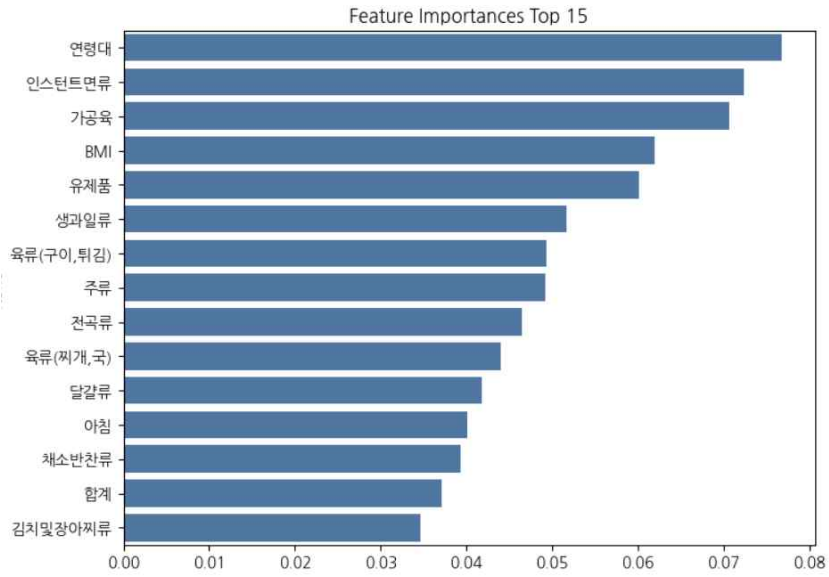
```
              precision    recall  f1-score   support

           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]]
```

모델링 결과 RandomForestClassifier(max_depth=12)일 때
F1_macro가 0.70이 나옴을 확인할 수 있다.



상위 15개의 Feature Importance를 확인해봤을 때, 실제 시각화에서도 가장 극명하게 차이가 드러났던 연령대가 가장 높은 중요도를 보인 것으로 확인되었다.

III. 결론

모델링을 통해
- 건강 이상 여부에서는 연령대가 가장 중요한 영향을 줌을 확인하였다.

시각화를 통해
- 식습관(음식 종류 섭취 빈도)에서는 너무 많은 빈도나 너무 적은 빈도가 아닌 때 가장 건강 이상 빈도가 낮음을 확인하였다.
- 식사 횟수가 너무 적거나, 채식 위주이거나, 저체중 또는 비만인 경우에는 그렇지 않을 때보다 건강 이상 위험이 더 높게 나왔음을 확인하였다.

이를 바탕으로 고연령자를 중점적으로 채식주의자, 식사 횟수가 너무 적은 사람, 저체중 또는 비만인 사람을 모니터링하면 국민 건강 증진에 도움을 줄 수 있을 것으로 판단된다.

다만 다음과 같은 한계점이 존재한다.
- 개인의 건강 응답에 주관이 들어가 있으므로 응답에 편향이 있을 수 있다.
- 모델링 결과 0에 대한 F1 score는 0.93인 반면, 1에 대한 F1 score는 0.47로 1을 상대적으로 잘 감지하지 못했다.