심장건강과 환자 변수 데이터분석을 통한 심장마비 위험 예측

3기 구혜준

프로젝트 개요

•데이터 선정 이유 / 분석목적 :

- 현대의학의 발전에도 불구하고 꾸준히 유지되는 심장마비 발생건수

  → 심장마비가 아직 해결하기 어려운 질환이라는 것을 의미

- 환자별 건강세부정보, 생활습관, 나라, 임금 등의 변수와 심장건강을 비교분석하여

  유의미한 결론을 도출해 심장마비 예방안을 모색

- Todo : 각종 변수, 또 파생변수가 heart attack risk 에 미치는 영향 살펴보기!

데이터셋 설명

- **데이터 출처** : 캐글 heart-attack-prediction-dataset

https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

- 주요 변수 설명
  - Cholesterol : 사람들의 콜레스테롤 수치
  - Exercise Hours Per Week : 사람들의 1주당 운동 시간
  - Income : 사람들의 수입
  - Heart Attack Risk : 심장마비 리스크 (0,1로 정의)
  - blood pressure을 최고치와 최저치로 정수화시킴
    - Systolic_BP : 최고치
    - Diastolic_BP : 최저치

데이터 전처리

- 라벨 인코딩

선택이유 : 일부 열의 데이터 타입이 숫자가 아닌 OBJECT(문자형) 으로 저장되어 있어 범주형 데이터 변환을 해야 한다.

Blood pressure은 최고치와 최저치로 분리하는 것이 데이터 분석에 편리

결과

Systolic_BP	Diastolic_BP
158	88
165	93
174	99
163	100
91	88

- 데이터 인코딩

선택 이유 : 모두 정수형 변수로 만들어 데이터 분석을 편리하게 만들기 위해

결과 요약 : 범주형 데이터에 숫자를 라벨링하고 float 값을 반올림하기 위해

```
# float_var 리스트형식으로 선언
float_var = ['Exercise Hours Per Week', 'Sedentary Hours Per Day', 'BMI']

# 원본 데이터를 별도로 저장
original_df = df[float_var].copy()

# 1의 자리까지 반올림 (정수 부분에서 반올림)
df[float_var] = df[float_var].round().astype(int)  # 반올림 후 정수형으로 변환

print(df.head())  # df 전체를 출력하면 반올림된 값 확인 가능
```

데이터 전처리

-기본적인 결측치 처리

선택이유 : 결측치를 제거하여 데이터 분석에 오류가 생기지 않도록

결과 : 완전한 데이터였기에 이상치는 존재하지 않음 -> 문구 띄웠음

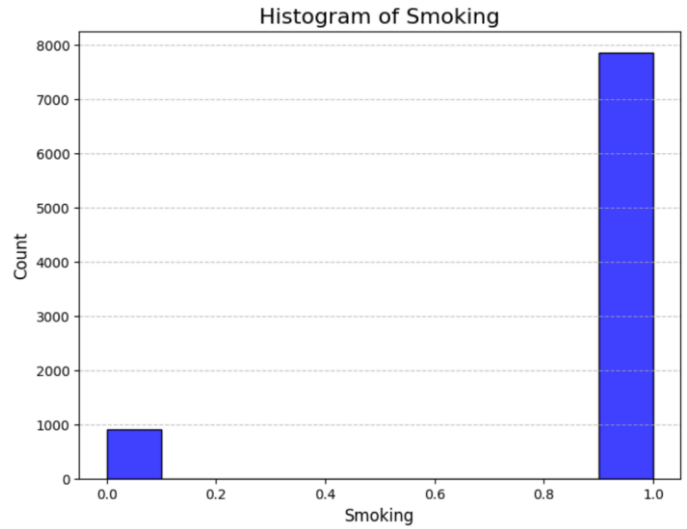
```
Systolic_BP                    0
Diastolic_BP                   0
dtype: int64
결측치가 없습니다. 결측치 처리 단계를 생략합니다.
```

- 데이터 인코딩

선택 이유 : 모두 정수형 변수로 만들어 데이터 분석을 편리하게 만들기 위해

결과 요약 : 범주형 데이터에 숫자를 라벨링하였고, float 값을 반올림하여 정수값으로 만듦

데이터 전처리



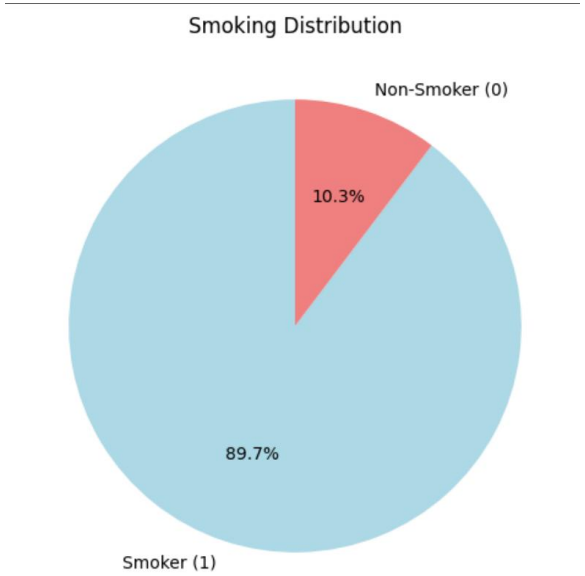
-이상치 처리

선택이유 : 유난히 삐져나온 값이 있으면 변수비교와 데이터분석에 잘못된 결과를 불러올 수 있기에

결과 : smoking 컬럼의 이상치 개수가 904개나 나옴

1. 0과 1이 아닌 값이 있나 확인하였고 2. 어떤 식으로 이상치가 존재하는지 시각화로 확인했음

→ 그냥 흡연자가 더 많은 편향된 데이터였다. (데이터 수집 과정이 편향되지 않았다고 가정, 그대로 진행)

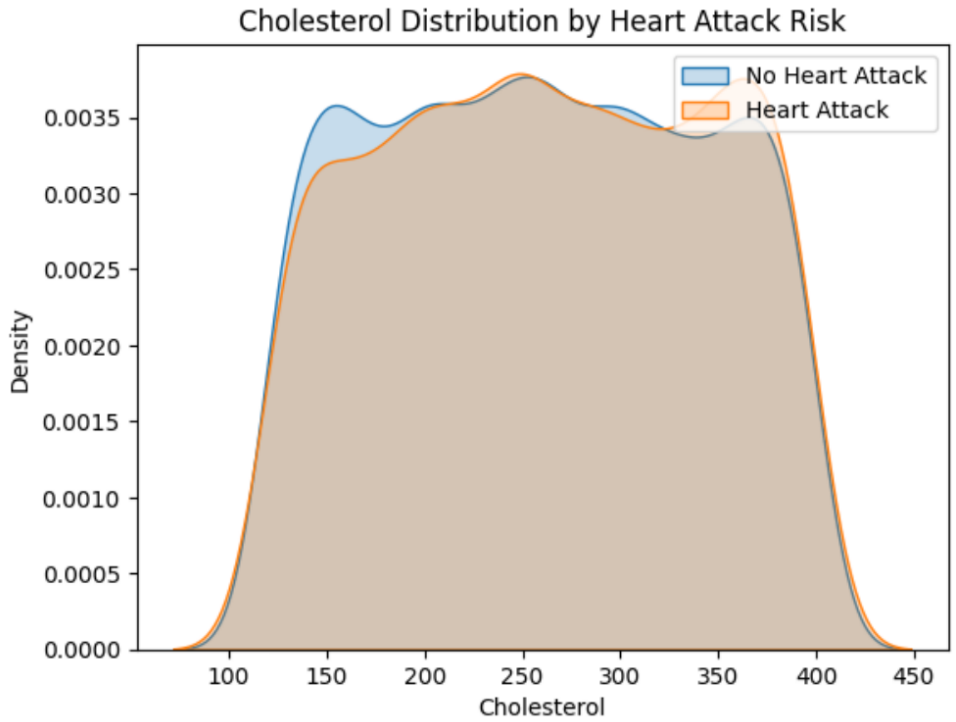



```
# 'Smoking' 컬럼의 고유 값 확인
print("Smoking 컬럼의 고유 값:", df['Smoking'].unique())

# 0과 1이 아닌 값의 개수 확인
invalid_values = df[~df['Smoking'].isin([0, 1])]
print("0과 1 이외 값의 개수:", len(invalid_values))
```

```
Smoking 컬럼의 고유 값: [1 0]
0과 1 이외 값의 개수: 0
```

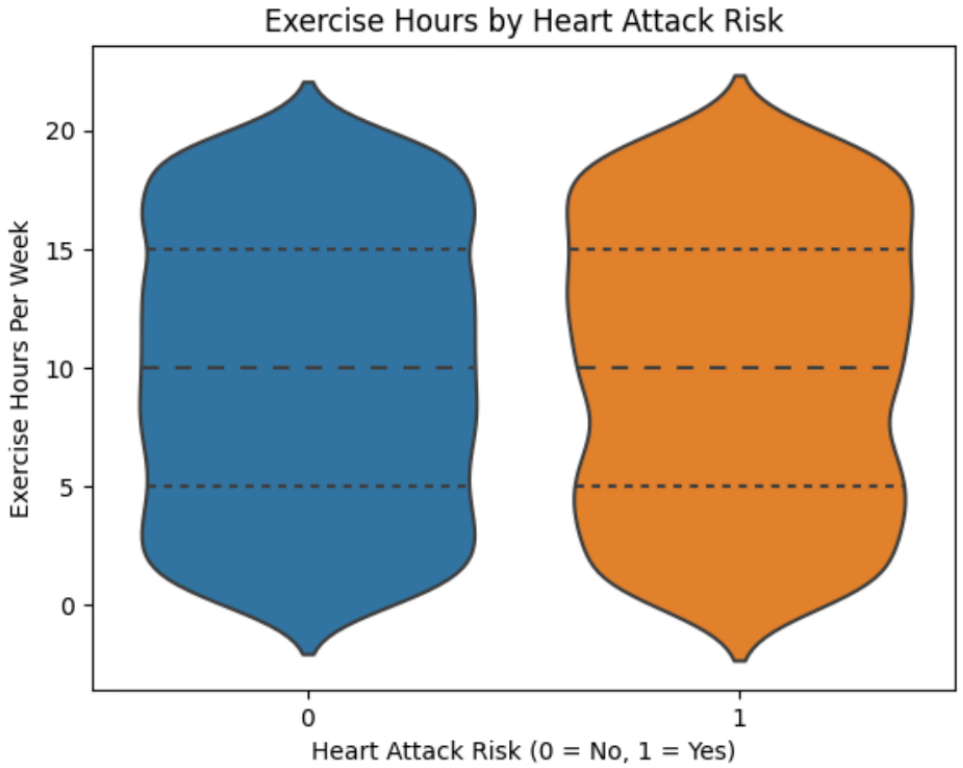
EDA 요약 (1)



- No heart attack 집단의 콜레스테롤 수치의 밀집도가 왼쪽에 좀 더 몰려있고, 반대로 heart attack 집단의 밀집도는 오른쪽에 몰려있음

→ 콜레스테롤 수치가 높을수록 심장마비 risk 높음

EDA 요약 (2)



- Heart attack risk가 1인 그래프에서 주 5~10시간 운동 부분의 허리가 오목형태이고 나머지부분은 risk가 0인 그래프와 비슷

→ 주 5~10시간, 적당한 운동을 하지 않는 사람들이 상대적으로 심장마비 위험 높음(이라 가정해보았음)

Feature Engineering

- 사용 기법 :

  1. **Inactive Obesity (BMI / 운동시간) :** 상호작용 변수**(**두 개의 독립적인 변수(BMI, Exercise Hours)를 결합해서 의미 있는 관계나 시너지 효과를 포착)

  2. **Inactive Lifestyle Index (좌식 시간 - 운동/7)** : 수치형 변수 조합 / 지표 생성- 두 개 이상의 연속형 변수 간의 연산으로 새로운 지표(지수)를 만든 것.

  3. **Hypertension Flag (혈압 기준 이진화) :** 연속형 변수(Systolic_BP)를 기준값으로 나눠 0/1로 분류

  4. **Income Level & Income Risk Interaction :** 구간화 (Binning): 소득을 분위수 기준으로 Low/Middle/High로 나눈 것 - 가중 위험도 : Income * Risk 로 만듦

  → 소규모 데이터셋에서 모델 성능을 크게 개선

결과 요약 및 인사이트

- **주요 분석결과 :**
    - 콜레스테롤 수치 증가 -> 심장마비 위험 증가
    - 소득 높을수록 심장마비 위험 증가
    - 고혈압이 있는 경우 심장마비 위험 증가
- **인사이트** :
    - 의외로 소득이 높을수록 심장마비 위험 증가
    - 아마 소득이 높을수록 고기와 기름진 음식을 많이 섭취하여
    - 고혈압과 콜레스테롤을 불러와 심장마비 위험이 증가했을수도?

현재까지의 한계점 : 상관관계를 매우 확실히 나타내주는 변수나 파생변수 발견 못함

향후 계획 및 개선 방향

- **추가 분석 아이디어** : 분석해보지 않은 모든 변수를 분석해본다. (특히 지역별 심장 리스크가 궁금함)
- **데이터 보완 계획** : 변수간 상관관계가 명확하지 않으므로, 다른 보충 데이터를 확보하여 현 데이터와 함께 적용하여 분석하고자 한다.
    - Ex) 소득수준과 국가별 경제수준(GDP데이터) 을 결합

        -> 국가별 개인의 상대 소득에 따라 소득별 심장리스크 분석

마무리

- Python도 처음이고, 데이터 분석도 처음
- 직접 분석을 통해 데이터를 내가 원하는 대로 정리하는 과정이 매력적!
- 하루종일 데이터분석만 해보고 싶다! 그래서 Fever day가 기대되는 바.