EDA 인사이트 보고서

건강검진 데이터를 활용한 흡연 패턴 분석

데이터 개요

7,000

전체 데이터

분석 대상 행(row) 수

18

변수 개수

전체 열(column) 수

0

결측치

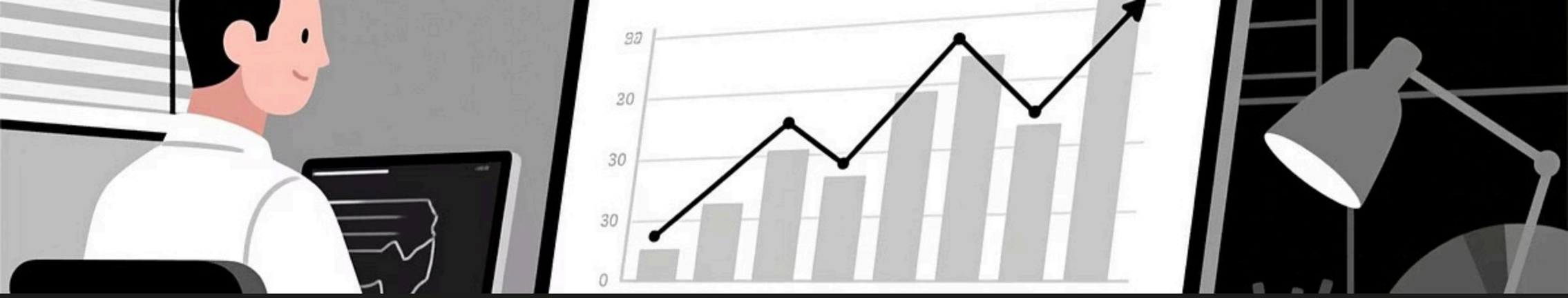
데이터 품질 우수

데이터 구성

신체 정보(나이, 키, 몸무게, BMI, 시력), 혈액 검사 값, 간 기능 지표, 콜레스테롤 수치, 단백질 수치 등 다양한 건강 지표를 포함합니다.

목표 변수: 흡연 여부 (0=비흡연, 1=흡연)

데이터 품질은 매우 양호하며, 기본적인 전처리 부담이 거의 없는 편입니다.



변수별 기본 통계 인사이트

나이

평균 **43.9세**, IQR 기준 35~50세가 가장 많습니다. 20대~30대에서 흡연자(label=1)의 비중이 상대적으로 높게 나타납니다.

체형 지표

키 평균 **164.8cm**, 몸무게 평균 **65.9kg**, BMI 평균 **24.1** (정상~과체중 경계). BMI가 높을수록 흡연자 비율이 증가하는 경향 확인 (상관계수 0.13).

공복혈당

평균 공복혈당 **99**, 75% 지점이 104, 최대치가 386으로 이상치 존재. 이상치 처리를 고려해야 할 변수입니다.

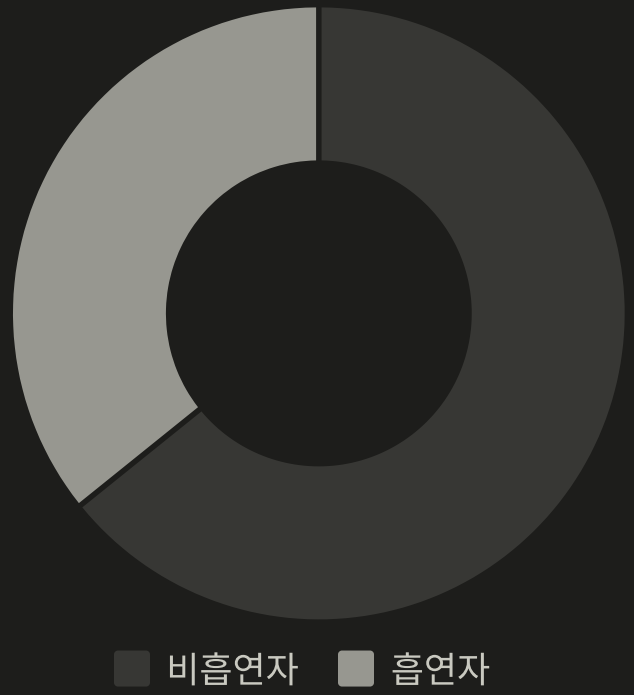
중성지방 & LDL

두 변수 모두 오른쪽으로 긴 꼬리를 가진 강한 오른쪽 편향 분포. 극단적으로 높은 값들이 다수 존재하여 로그 변환 또는 winsorization 고려가 필요합니다.

시력 & 총치

대부분 분포가 0~1 사이에 몰려 있어 모델 영향력이 낮을 가능성이 높습니다.

흡연 여부 분포 분석



클래스 불균형 확인

비흡연자 **4,500명 이상**, 흡연자 **2,500명 정도**로 **약 65:35 비율**의 클래스 불균형이 존재합니다.

모델링 단계에서 SMOTE, class_weight 적용 등의 불균형 해소 기법이 필요할 수 있습니다.

나이별 흡연 패턴

20~40대

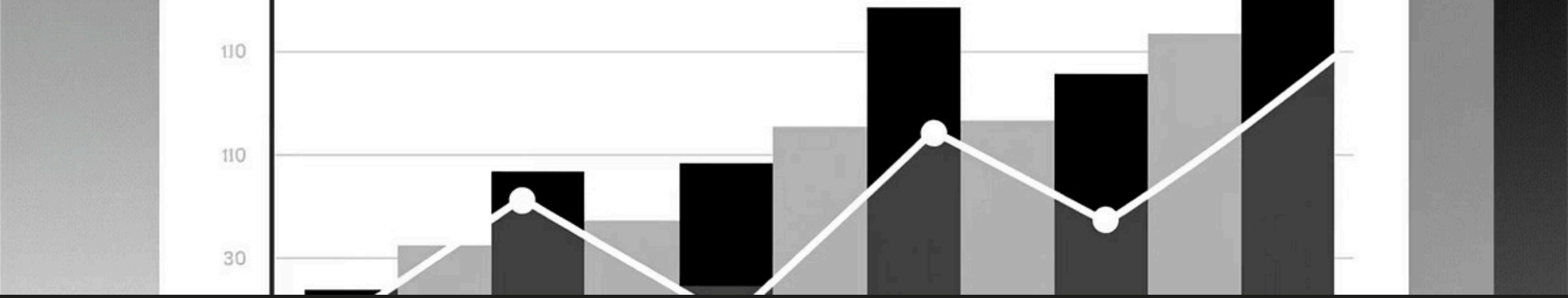
흡연자(label=1) 비중이 확실히 높은 연령대

50대 이후

흡연자 비율이 급격히 감소하는 구간

> "연령대에 따른 흡연 패턴 변화"는 모델링에 중요한 피처로 작용할 가능성이 높습니다.

젊은 연령층에서 흡연율이 높다가 나이가 들면서 건강 인식이 높아지거나 금연하는 경향이 뚜렷하게 나타납니다. 이는 연령을 예측 모델의 핵심 변수로 활용할 수 있는 근거가 됩니다.



변수별 분포 분석 핵심 인사이트

전반적인 특징

다수의 변수들이 오른쪽 꼬리가 긴 분포(Right-skewed)를 보이며, 일부 변수에서는 극단값(outlier)이 존재합니다. 분포가 경직된 변수(시력, 총치 등)는 정보량이 적어 모델 영향이 적을 가능성이 있습니다.

전처리 필요 변수

공복혈당, 중성지방, LDL, 콜레스테롤, 저밀도지단백 등은 극단값이 매우 많아 변환 또는 이상치 처리가 필수적입니다.

안정적 피처

BMI, 몸무게, 키는 비교적 정규분포에 가까워 안정적인 피처로 활용 가능합니다.

상관관계 분석 핵심 인사이트

흡연(label)과 상관 높은 변수

39%

키(cm)

양의 상관관계

31%

몸무게(kg)

양의 상관관계

29%

LDL

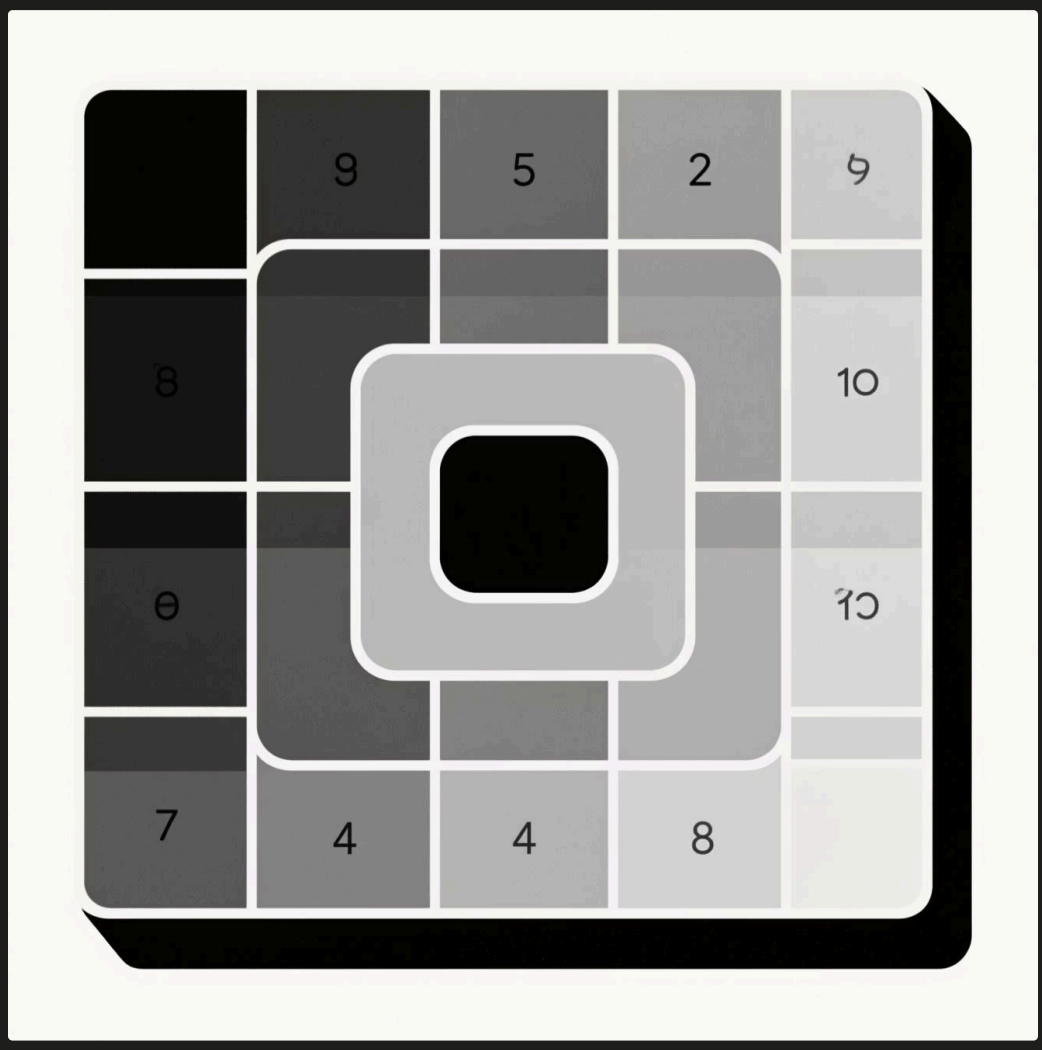
저밀도지단백

25%

중성지방

양의 상관관계

HDL(고밀도지단백): -0.18의 음의 상관. HDL이 낮을수록 흡연자일 가능성이 높습니다.



흡연자일수록 체중이 높고, LDL과 중성지방이 높아지는 패턴이 확인됩니다. 이는 실제로 흡연이 대사 건강에 악영향을 주는 경향과 일치합니다.

변수 간 높은 상관 조합

- 몸무게 ↔️ 중성지방 (0.34)
- LDL ↔️ 총콜레스테롤 (0.72) - 다중공선성 위험