

# 생체 신호 기반 흡연 여부 예측 문제에서 본 생성 데이터의 한계와 통찰

[C4] Smokalyzer

2025.06.04



고주리

- Github
- Modeling
- Feature Engineering
- Model Tuning

김귀연

- 발표
- Feature Engineering
- Model Tuning

문선영

- EDA
- 데이터 전처리
- 발표 자료 정리

채승희

- 데이터 분석
- 시각화
- Feature Engineering
- 발표 자료 작성

# 목차

## 데이터 소개

01

- 1. 데이터 개요
- 2. 주요 변수
- 3. 외부 데이터셋



## 분석 목적 및 전략

02



## 프로젝트 프로세스

03



## 탐색적 데이터 분석(EDA)

04

- 1. 타겟변수 smoking
- 2. 변수 분포 시각화
- 3. 피처 간 상관관계 (다중공선성)



## 데이터 전처리 및 가공

05

- 1. 결측치 확인
- 2. 이상치 탐지 및 처리
- 3. Feature Engineering



## 모델링과 성능 평가

06

- 1. 모델 선정
- 2. 베이스라인 모델 성능 비교
- 3. Feature Selection
- 4. 양상별 모델 비교
- 5. 최종 모델 선정
- (6. 이슈 트래킹)



## 프로젝트 인사이트

07

- 1. 시도 대비 결과
- 2. 결과 분석
- 3. 하이퍼파라미터 관련 차이점 분석
- 4. 결론 및 개선 방향
- 5. 분석 과정에서 얻은 인사이트



# 1. 데이터 소개

## 1-1. 데이터 개요

### Binary Prediction of Smoker Status using Bio-Signals

생체 신호를 이용한 흡연자 상태의 이진 예측

#### 사회적 배경

- WHO(세계보건기구) 보고에 따르면 2030년까지 흡연으로 인한 사망자 수가 1,000만 명에 이를 것으로 예측
- 금연을 위한 다양한 치료법이 제시되었지만, 성공률이 낮으며 많은 의사들이 금연 상담을 비효과적이라 여김  
→ 흡연 가능성 및 금연 성공 확률을 예측할 수 있는 머신러닝 모델에 주목

#### 과제 개요

- 바이오 시그널 데이터를 기반으로 개인의 흡연 여부(smoking)를 예측
- 실제 데이터를 기반으로 생성된 합성 데이터(synthetic data) 사용
- 예측 방식
  - 이진 분류
  - 제출 결과: 확률 값
- 평가 지표: ROC-AUC

#### 데이터 구성

- train.csv**
  - 학습 데이터셋 (smoking 포함)
  - 개수: 159256, 24
- test.csv**
  - 테스트 데이터셋 (smoking 제외)
  - 개수: 106171, 23
- Target 변수: **smoking**

#### 예측 모델의 기대 효과

- 건강검진을 통한 고위험군 조기 선별
- 의료진의 상담 효율성 향상
- 흡연이 영향을 미치는 생체 신호 탐지

# 1. 데이터 소개

## 1-2. 주요 변수

구분	변수명	설명
신체 정보	height, weight, waist, age	기본 체형, 나이
시각/청각	eyesight(left/right), hearing(left/right)	시력, 청력 검사 결과
혈압/혈당	systolic, relaxation, fasting blood sugar	고혈압, 당뇨 여부 반영
지질 지표	cholesterol, triglyceride, HDL, LDL	이상지질혈증 관련 수치
간 기능 지표	AST, ALT, Gtp	흡연·음주로 영향받는 대표 지표
기타 건강지표	hemoglobin, serum creatinine, urine protein, dental caries	혈액, 신장, 구강 관련 정보

### 간기능검사

#### ▶ 목적

간세포가 손상 받는 경우 간세포 내에 존재하는 효소들이 혈중으로 방출되어 혈중수치가 증가하게 되고 이로 인해 간의 기능 이상을 진단하게 됩니다.

#### ▶ 참고치

- AST(GOT)/ALT(GPT) : 0 - 40 IU/L
- r-GTP : 남자 11 - 63 IU/L, 여자 8 - 35 IU/L
- 알칼라인 포스파타제 (ALP) : 40 - 120 IU/L
- 총빌리루빈 (Total Bilirubin) : 0.2 - 1.2 mg/dL
- 총단백 (Total Protein) : 6.0-8.0 g/dL
- 알부민 (Albumin) : 3.3 - 5.2 g/dL

#### ▶ 참고치

혈압분류	수축기혈압(mmHg)	이완기혈압(mmHg)
정상혈압	<120	그리고 <80
주의혈압	120~129	그리고 <80
고혈압전단계	130~139	또는 80~89
고혈압	1기 140~159 2기 ≥160	또는 90~99 ≥100
수축기단독고혈압	≥140	그리고 <90

대한고혈압학회, 2018년 고혈압 진료지침

서울아산병원 건강증진센터

# 1. 데이터 소개

## 1-3. 외부 데이터셋

### Body Signal of smoking

#### 외부 데이터셋 사용 이유

- 주 데이터셋에는 존재하지 않는 gender 피처를 생성하기 위해
- 외부 데이터로부터 gender 예측 모델을 학습한 뒤 → 주 데이터에 남성일 확률(gender)을 새로운 feature로 추가
- 확률값(0~1)으로 표현하여 단순한 이진 분류 예측 결과보다 애매하거나 불확실한 결과 보완

#### 데이터셋 비교

구분	주 사용 데이터셋	외부 데이터셋
출처	Kaggle Competition 용도 합성 데이터	Kaggle 생체 신호 데이터 모음
샘플 수	(159,256, 24)	(55,692, 27)
공통 피처	23개 건강 생체 지표 (age, height, cholesterol 등)	
차별 피처	gender 없음	gender, oral, tartar 포함
목적	흡연 여부 예측	흡연 여부 예측 → 우리는 성별을 예측하는 모델로 사용



#### Body signal of smoking

find smokers by vital signs (binary classification)

[kaggle.com](https://www.kaggle.com)

두 데이터 셋의 구조와 목적 유사!

외부 데이터의 추가 정보

gender(성별)를  
확장 피처로 생성하여 사용

## 2. 분석 목적 및 전략



### 데이터 선정 동기/배경

- Smoker Status Prediction using Bio-Signals 데이터셋을 기반으로 딥러닝 모델을 통해 생성된 합성 데이터
- 이진 분류에 최적화되어 다양한 분류 ML 알고리즘 적용과 성능 비교에 용이
- 데이터의 규모와 복잡도가 균형적이고 적절하여 전처리부터 모델링까지 데이터 분석의 전 과정 실습에 적합하다고 판단



### 분석 목적

- 바이오 신호를 기반으로 생성 데이터를 가지고 흡연 여부 예측 시, ROC-AUC 점수를 높일 수 있는 방안



### 접근 전략

- 도메인 기반 파생 피쳐생성
- 다중공선성 제거 기반 피쳐 선택
- 개별 모델의 하이퍼파라미터 튜닝 및 soft voting 양상을



### 도메인 해석

- 서울아산병원 건강증진센터
- Changes in Oxygen Saturation and Walk in Relation to Smoking and Types of Shoes 와 같은 건강에 대한 공식 지표 및 연구 사례를 바탕으로 데이터를 해석하고 이상치 처리, 상관성 분석



### 협업 방안

- 코드 공동 작업 : Colab + Google Drive
- 프로젝트 관리 및 버전 관리 : GitHub Organization
- 발표자료 공동 제작: Canva

## Meetings

Keep your discussions list and action items neatly organized inside meeting notes.

- 데이터 후보
- Reference
- Check List
- GitHub Organization 사용 가이드

### All meetings

Aa Name	Date	안건	출결사항
💡 Brainstorm Session	2025년 5월 29일 오후 4:30	1. 팀 이름 정하기 2. 데이터 셋 정하기 (~4시)	김민석님 휴가, 문선영님 외출
🚀 Day1	2025년 5월 30일	데이터 전처리 및 모델링 학습	
👉 Day2	2025년 6월 2일	결과 비교 및 최대 성능 도출	문선영님 휴가
📄 Day3	2025년 6월 3일	결과 정리 및 ppt 작성	
📝 Day4	2025년 6월 4일 오후 2:30	- 발표 시간 : 팀당 최대 30분(발표 15분 + QnA 및 피드백 15분)	

## ◀ 회의록 작성

공유 문서함 > Smokalyzer

유형 ▾ 사람 ▾ 수정 날짜 ▾ 출처 ▾

이름	소유자	내가 마...	파일 크기	⋮
Original	kami.gy.kim	2025. 5. 30.	—	⋮
v_4.0_smokalyzer_final_notebook.ipynb	나	오전 10:07	5.7MB	⋮
v_3.0_smokalyzer_feature_engineering.ip...	나	오전 10:02	5.8MB	⋮
v_2.0_smokalyzer_feature_engineering_고...	나	오전 7:17	5.7MB	⋮
v_2.0_smokalyzer_feature_engineering_고...	나	2025. 6. 3.	5.7MB	⋮
v_1.0_smokalyzer_baseline.ipynb	나	2025. 6. 3.	1.7MB	⋮

## Project Progress

## ◀ 프로젝트 진행 상황 기록

▶ 프로젝트 진행 상황 및 에러 해결 과정 기록

▣ 칸반 보드

☰ 표 보기

…

새로 만들기

### ✓ Progress

○ 할 일 13

▷ 진행 중 15

☑ 완료 10

▼ 전처리 14

작업 내용 (예시)

☒ 피쳐 분석

선영

☒ 코드 리팩토링

주의

☒ 파생 피쳐 다중공선성 확인

주의

☒ Feature Selection 실험

주의

+ 새 페이지

☒ [EDA] 원본 - 피쳐 상관관계 분석

체 채승희

☒ 노트북 목차

Gwiyeon Kim

☒ BMI 피쳐 + 이상치 클리핑

주의

☒ 데이터 전처리 프로세스

주의

☒ 피쳐 생성: gender + BP\_level

체 채승희

☒ 중요 변수 선별(SHAP)

Gwiyeon Kim

☒ 피쳐 생성 실험: 혈액글로빈 중심 상호 작용 파생변수

주의

## ▲ Google Drive를 활용한 코드 공유 및 협업

## ▼ 버전 단위 수정사항 커밋

-o Commits on Jun 4, 2025

feat: v\_3.0 모델 성능 개선

jul-ee committed 1 hour ago

-o Commits on Jun 3, 2025

fix: v\_2.1 submission 생성 방식 수정

jul-ee committed 20 hours ago

feat: v\_2.0 피쳐 엔지니어링 적용

jul-ee committed yesterday

feat: v\_1.0 baseline 코드 노트북 추가

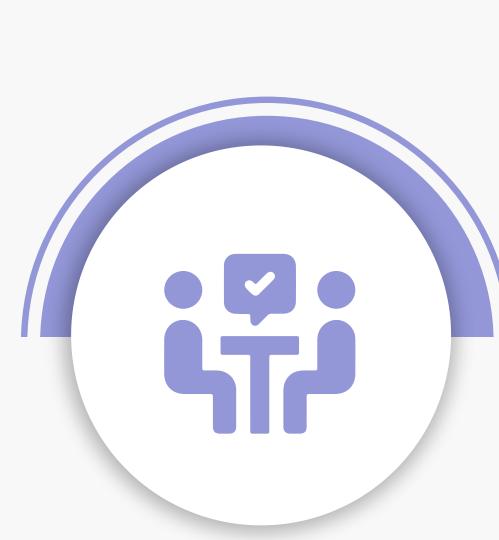
jul-ee committed yesterday

-o Commits on Jun 2, 2025

# 3. 프로젝트 프로세스

- 해결하고자 하는 문제 정의
- 이진 분류 문제
- 프로젝트의 최종 목표 설정

## 문제 정의 및 목표 설정



## 데이터 선정 및 팀 역할 분담

- 후보 중 분석 대상 데이터 확정
- 팀 내 역할 분배
  - 전처리, 모델링, 분석, 발표



## EDA(탐색적 데이터 분석)

- 변수 이해
- 분포 파악
- 이상치/결측치 확인
- 주요 인사이트 및 가설 수립



## 데이터 전처리



- 결측치 처리, 인코딩, 스케일링 등
- 모델에 투입 가능한 형태로 정제

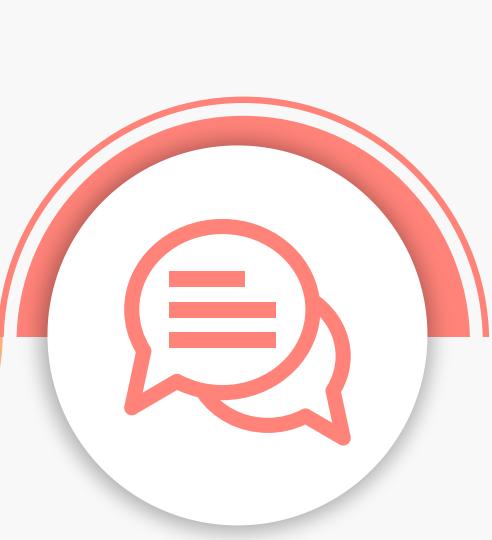
- 결과 해석, 변수 중요도 분석
- 그래프/표를 통해 핵심 전달

## 인사이트 도출 및 시각화



## 모델링 및 성능 평가

- 알고리즘 선택 및 학습
- 성능 평가 및 교차검증

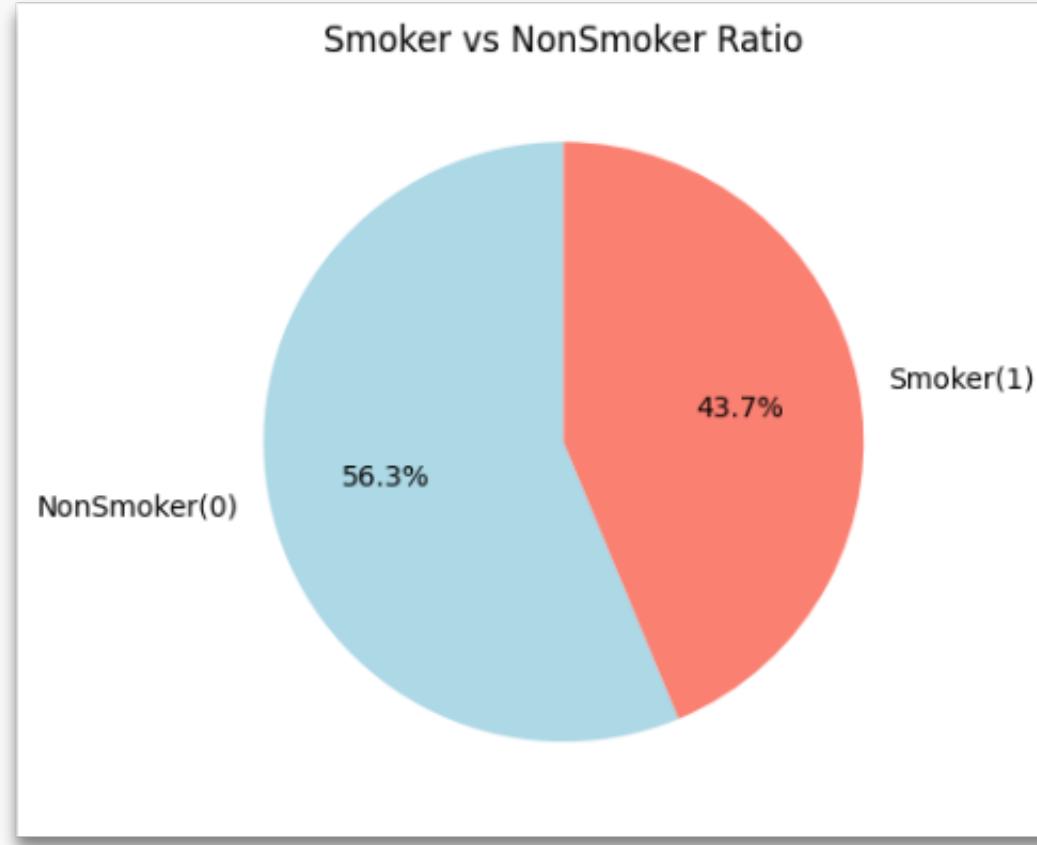


## 발표 및 회고

- 스토리 중심 발표 준비
- 피드백 및 프로젝트 회고 공유

# 4. 탐색적 데이터 분석(EDA)

## 4-1. 타겟 변수 SMOKING

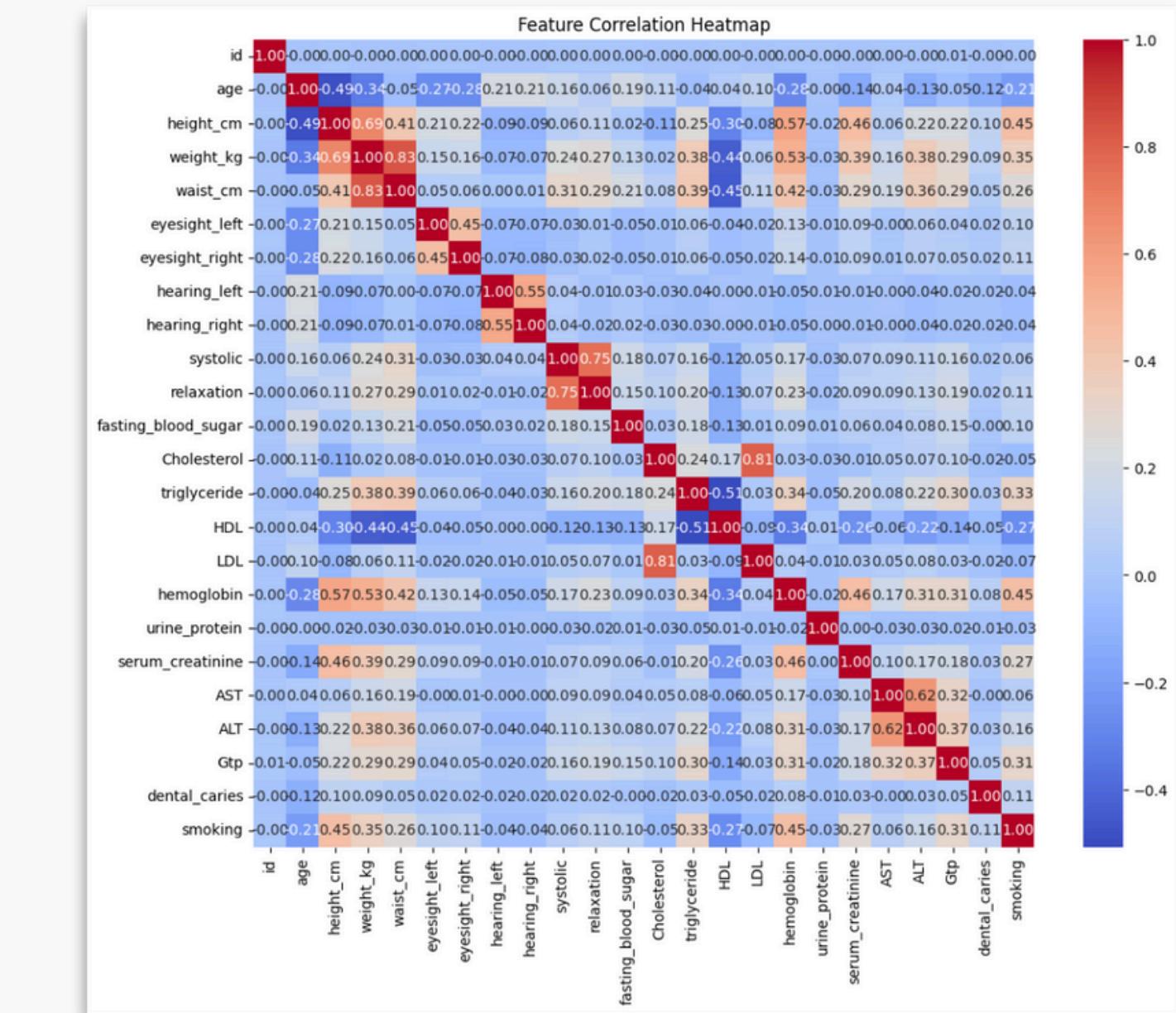


# 타겟 변수 분포

# 비흡연자 vs 흡연자

- 비흡연자(NonSmoker=0) 비율 56.3%
  - 흡연자(Smoker=1) 비율 43.7%

 클래스 비율은 비교적 균형에 가까운 분포



## 변수 간 상관관계 히트맵

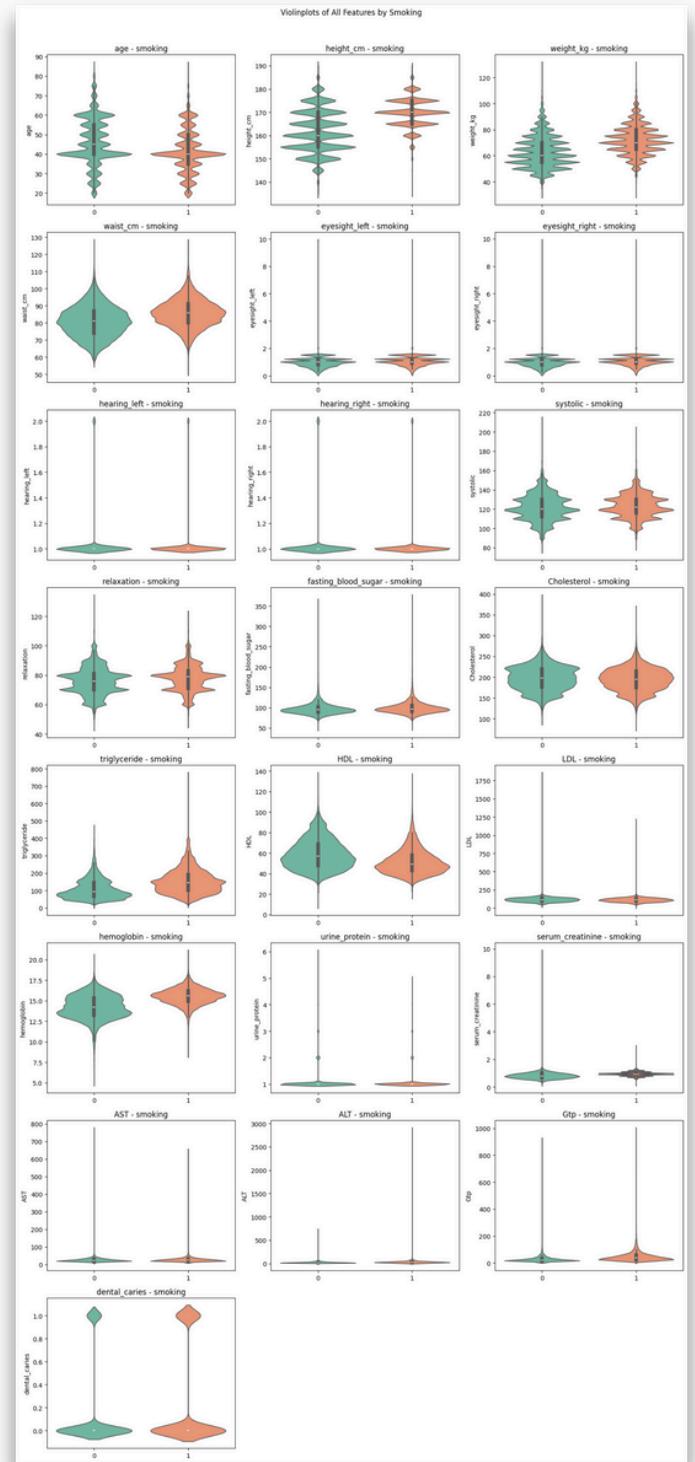
smoking과의 상관성 중심

1. hemoglobin +0.45
  2. height\_cm +0.45
  3. weight\_kg +0.35
  4. triglyceride +0.33
  5. waist\_cm +0.26
  6. HDL -0.27
  7. Gtp +0.26

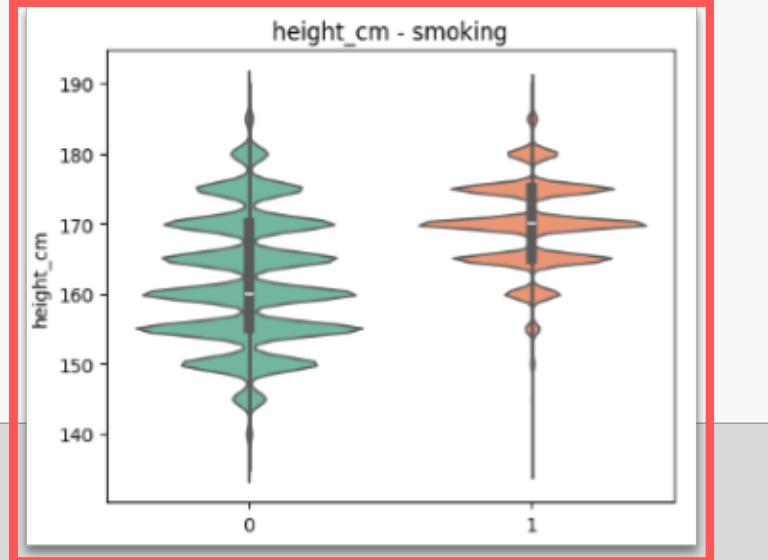
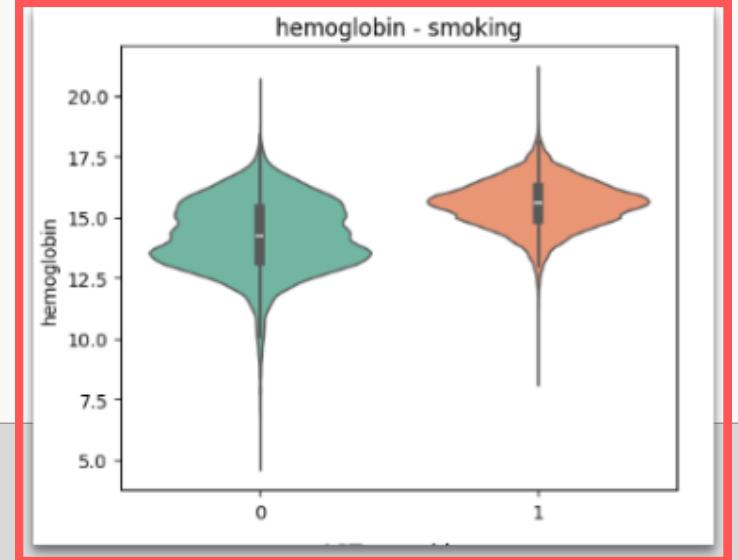
- ☞ 흡연 여부는 체형, 혈중 지질, 간 기능 등의 지표들과 유의한 상관관계를 보이나 생각보다 크게 유의미한 지표 X  
→ Feature Engineering 고려

# 4. 탐색적 데이터 분석(EDA)

## 4-2. 변수 분포 시각화



	피처	해석
hemoglobin		흡연자가 살짝 더 높은 중앙값을 가짐
height(cm)		흡연자 그룹이 전반적으로 키가 더 크고, 분포가 더 좁게 집중됨
weight(kg)		흡연자 쪽이 약간 더 무거움
waist(cm)		흡연자 쪽 허리둘레가 큼
systolic / relaxation (혈압)		흡연자 쪽이 중앙값도 높고 분포도 우측으로 더 퍼져 있음
Gtp / ALT / AST (간수치)		흡연자 쪽에 긴 꼬리(극단치)와 분포 중심 상승 → 간 기능 관련 이상 가능성
LDL / Cholesterol / triglyceride		흡연자가 높은 밀도 중심 + 더 많은 고지혈 이상치 존재
HDL		흡연자에서 좋은 콜레스테롤(HDL)이 더 낮음 → 일반적 생리학 패턴과 일치함
fasting blood sugar		흡연자 쪽이 약간 더 높고 이상치 더 많음, 하지만 완전히 구분되진 않음



# 4. 탐색적 데이터 분석(EDA)

## 4-3. 피처 간 상관관계(다중공선성)

feature	VIF
const	1244.431309
Cholesterol	7.333379
weight_kg	6.801537
LDL	6.107157
waist_cm	4.588011
HDL	3.221739
triglyceride	3.001240
height_cm	2.914232
systolic	2.456500
relaxation	2.409581
ALT	1.986468
hemoglobin	1.848829
age	1.819872
AST	1.742144
hearing_right	1.465960
hearing_left	1.464441
serum_creatinine	1.405493
Gtp	1.335092
eyesight_right	1.303434
eyesight_left	1.295456
fasting_blood_sugar	1.126582
dental_caries	1.021530
urine_protein	1.005057
id	1.000136

	Feature 1	Feature 2	Correlation
66	weight_kg	waist_cm	0.830208
212	Cholesterol	LDL	0.808533
171	systolic	relaxation	0.753003
45	height_cm	weight_kg	0.686645
266	AST	ALT	0.623408

다중공선성 VIF  
5 이상

- Cholesterol
- weight\_kg
- LDL

피처 간 상관성  
0.6 이상 확인

- weight(kg) - waist(cm)
- Cholesterol - LDL
- systolic - relaxion
- height(cm) - weight(kig)
- AST - ALT



Feature Engineering  
Feature Selection

# 5. 데이터 전처리

## 5-1. 결측치 확인

```
train_df.info() #결측치가 없음
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 159256 entries, 0 to 159255
Data columns (total 24 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   id               159256 non-null   int64  
 1   age              159256 non-null   int64  
 2   height(cm)       159256 non-null   int64  
 3   weight(kg)       159256 non-null   int64  
 4   waist(cm)        159256 non-null   float64 
 5   eyesight(left)   159256 non-null   float64 
 6   eyesight(right)  159256 non-null   float64 
 7   hearing(left)    159256 non-null   int64  
 8   hearing(right)   159256 non-null   int64  
 9   systolic          159256 non-null   int64  
 10  relaxation        159256 non-null   int64  
 11  fasting blood sugar 159256 non-null   int64  
 12  Cholesterol      159256 non-null   int64  
 13  triglyceride     159256 non-null   int64  
 14  HDL              159256 non-null   int64  
 15  LDL              159256 non-null   int64  
 16  hemoglobin       159256 non-null   float64 
 17  Urine protein    159256 non-null   int64  
 18  serum creatinine 159256 non-null   float64 
 19  AST              159256 non-null   int64  
 20  ALT              159256 non-null   int64  
 21  Gtp              159256 non-null   int64  
 22  dental caries    159256 non-null   int64  
 23  smoking           159256 non-null   int64  
dtypes: float64(5), int64(19)
memory usage: 29.2 MB
```

```
test_df.info() #결측치가 없음
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 106171 entries, 0 to 106170
Data columns (total 23 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   id               106171 non-null   int64  
 1   age              106171 non-null   int64  
 2   height_cm        106171 non-null   int64  
 3   weight_kg        106171 non-null   int64  
 4   waist_cm         106171 non-null   float64 
 5   eyesight_left   106171 non-null   float64 
 6   eyesight_right  106171 non-null   float64 
 7   hearing_left    106171 non-null   int64  
 8   hearing_right   106171 non-null   int64  
 9   systolic          106171 non-null   int64  
 10  relaxation        106171 non-null   int64  
 11  fasting_blood_sugar 106171 non-null   int64  
 12  Cholesterol      106171 non-null   int64  
 13  triglyceride     106171 non-null   int64  
 14  HDL              106171 non-null   int64  
 15  LDL              106171 non-null   int64  
 16  hemoglobin       106171 non-null   float64 
 17  urine_protein    106171 non-null   int64  
 18  serum_creatinine 106171 non-null   float64 
 19  AST              106171 non-null   int64  
 20  ALT              106171 non-null   int64  
 21  Gtp              106171 non-null   int64  
 22  dental_caries    106171 non-null   int64  
dtypes: float64(5), int64(18)
memory usage: 18.6 MB
```

# 5. 데이터 전처리

## 5-2. 이상치 탐지 및 처리

변수명	특징 및 도메인 해석	처리 방안
eyesight (좌/우)	시력 9.9는 존재 불가능한 값 → 데이터 오류로 판단	9.9 → 2.0으로 수정
triglyceride	최대 766 → 중성지방 수치 높음(의학적으로 가능)	수치 유지, 제거 없음
LDL	300 이상은 생리적으로 매우 드문 값 → 데이터 오류로 판단	300 초과 → 300으로 클리핑
hemoglobin	hemoglobin은 21로 높은 값 가능	수치 유지, 제거 없음
serum creatinine	serum creatinine는 9.9로 높은 값 가능	수치 유지, 제거 없음
ALT	독성 간염, 약물 유발성 손상으로 1,000 U/L 까지는 가능 → 초과는 데이터 오류로 판단	1,000 초과 → 1,000으로 클리핑
Gtp	간질환, 알코올성 질환, 종양, 폐쇄 등 다양한 요인 존재 → 500~999 수치 가능	수치 유지, 제거 없음

# 5. 데이터 전처리

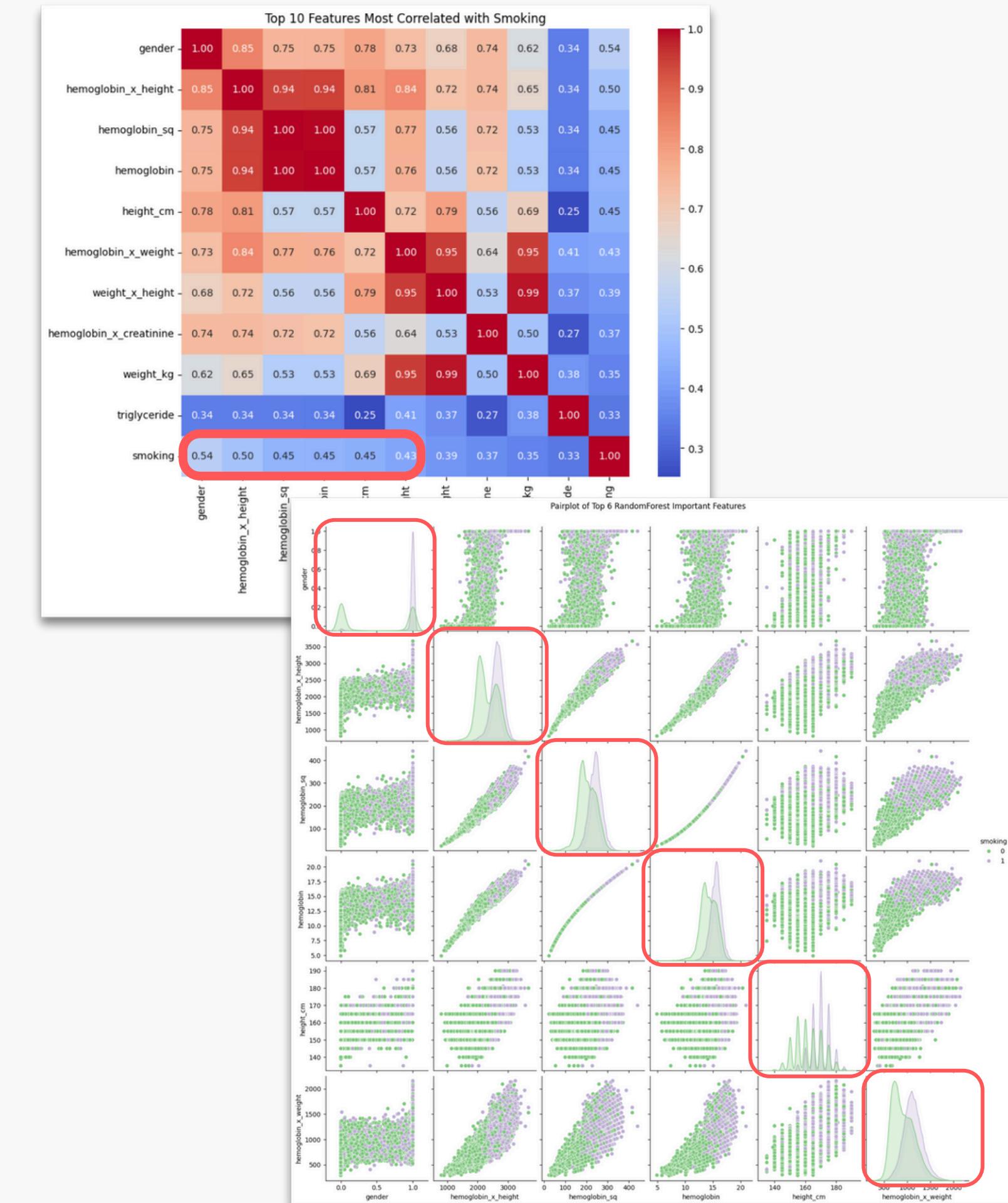
## 5-3. FEATURE ENGINEERING (1/2)



# 5. 데이터 전처리

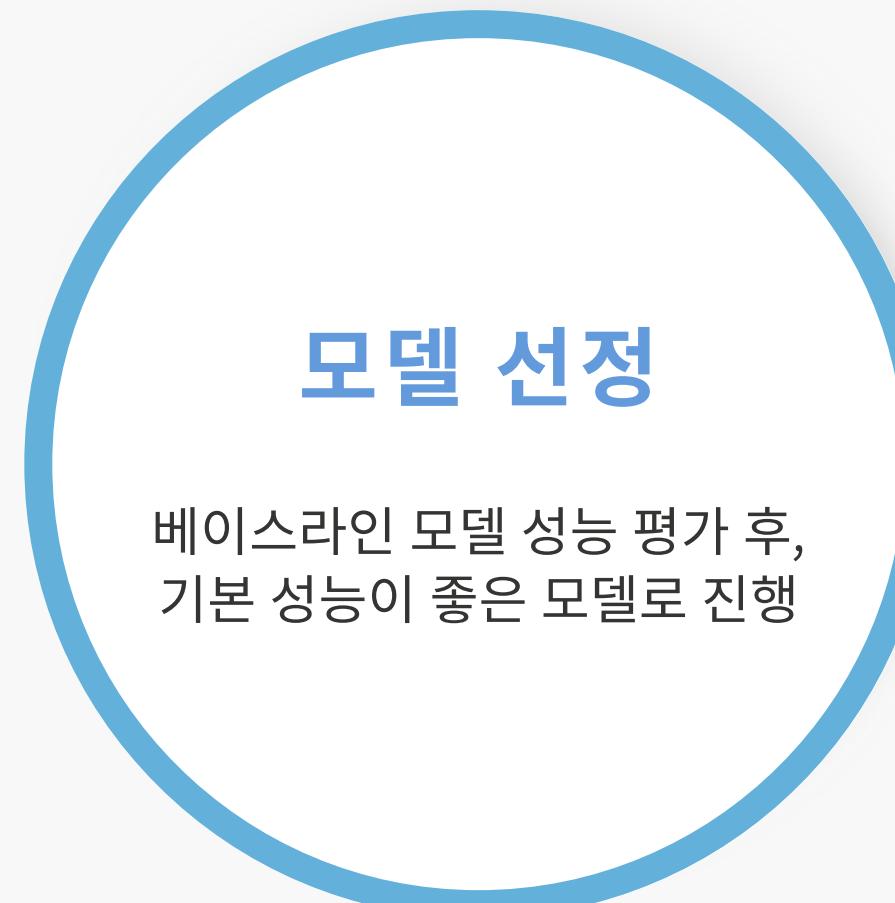
## 5-3. FEATURE ENGINEERING (2/2) - 시각화

순위	변수명	상관계수	해석
1	gender	0.54	성별에 따라 흡연 여부 큰 차이 존재. <u>남성일수록 흡연 비율 높음</u>
2	hemoglobin × height	0.50	산소 운반능력 × 체격 → 흡연자의 생리적 보상 반응과 체형 반영
3	hemoglobin_sq	0.45	혈색소 비선형 효과 반영, 흡연 시 수치 증가 경향
4	hemoglobin	0.45	<u>흡연자는 산소 부족 보상</u> 을 위해 혈색소 수치 상승 가능
5	height_cm	0.45	<u>성별 간 체형 차이 반영</u> 가능, 흡연자에서 상대적으로 신장이 큰 경향
6	hemoglobin × weight	0.43	체중과 혈색소의 결합 → 건강 상태 및 대사 요인 반영 가능성



# 6. 모델링과 성능 평가

## 6-1. 모델 선정



# 6. 모델링과 성능 평가

## 6-2. 베이스라인 모델 성능 비교



### 베이스라인 모델

이상치만 처리하고, 파생 변수(Feature Engineering을 통해 생성한 결과)를 모두 포함하여, 기본적인 파라미터 설정으로 학습한 모델  
→ 각 모델의 기본 분류 성능을 비교하고, 향후 튜닝 및 양상을 전략의 출발점

Model	Accuracy	Precision	Recall	F1	ROC-AUC	
CatBoost	0.7818	0.7203	0.8193	0.7666	0.8659	<b>모델 활용 방향 수립</b>
LightGBM	0.7798	0.7195	0.8137	0.7637	0.8644	<ul style="list-style-type: none"><li>CatBoost, LightGBM과 XGBoost가 전반적으로 가장 우수한 성능을 보임 → 이후 주요 실험과 하이퍼파라미터 튜닝에 중심적으로 활용</li></ul>
XGBoost	0.7777	0.7184	0.8089	0.7610	0.8643	<ul style="list-style-type: none"><li>다른 모델들은 양상을 전략(Voting, Stacking) 테스트에 다양성을 실험하기 위한 보조 모델로 활용</li></ul>
RandomForest	0.7717	0.7108	0.8060	0.7554	0.8561	
AdaBoost	0.7662	0.6998	0.8149	0.7530	0.8483	<ul style="list-style-type: none"><li>ROC-AUC 기준으로 0.86 이상의 성능 확보 → 향후 모델 조합 및 튜닝에서 기준 성능 벤치마크</li></ul>
LogisticRegression	0.7601	0.6903	0.8187	0.7491	0.8426	

# 6. 모델링과 성능 평가

## 6-3. FEATURE SELECTION (1/2)

### 1. Feature Selection 목적

- 다중공선성 문제 해소 및 예측 성능 향상
- 해석력 개선과 일반화 성능 확보

### 2. 제거 기준

기준	설명	제거 기준
VIF (다중공선성)	100 이상 → 강한 다중공선성 우려	제거
타겟 상관계수	0.1 미만 → 타겟 예측에 기여도 낮음	제거
Feature Importance	0.01 미만 → RandomForest 모델 내 기여 거의 없음	제거

### 3. 실제 효과

- 베이스라인 모델에 적용  
→ 대부분의 모델 성능 하락

☞ 추가적인 피처 선별 과정이 필요!

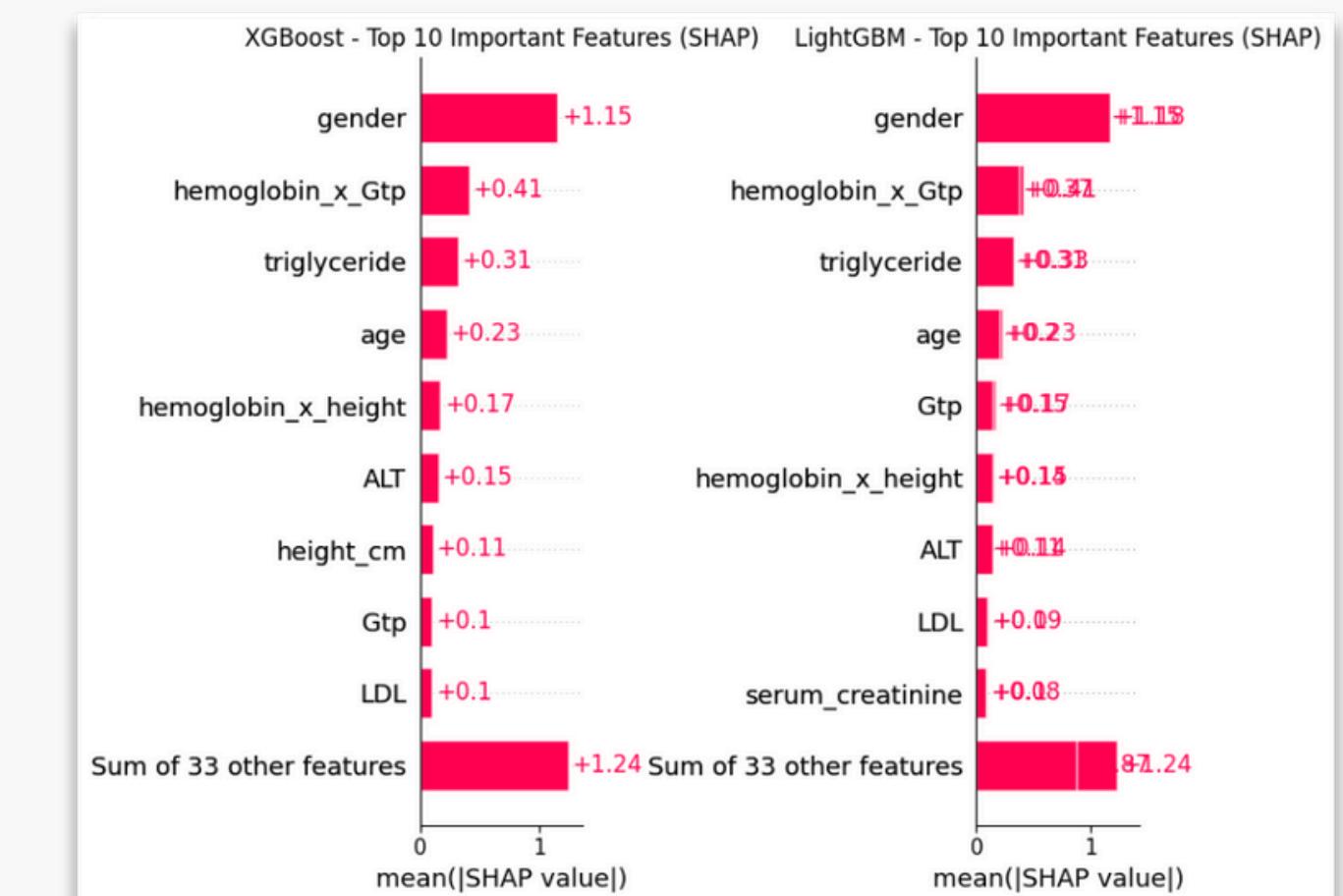
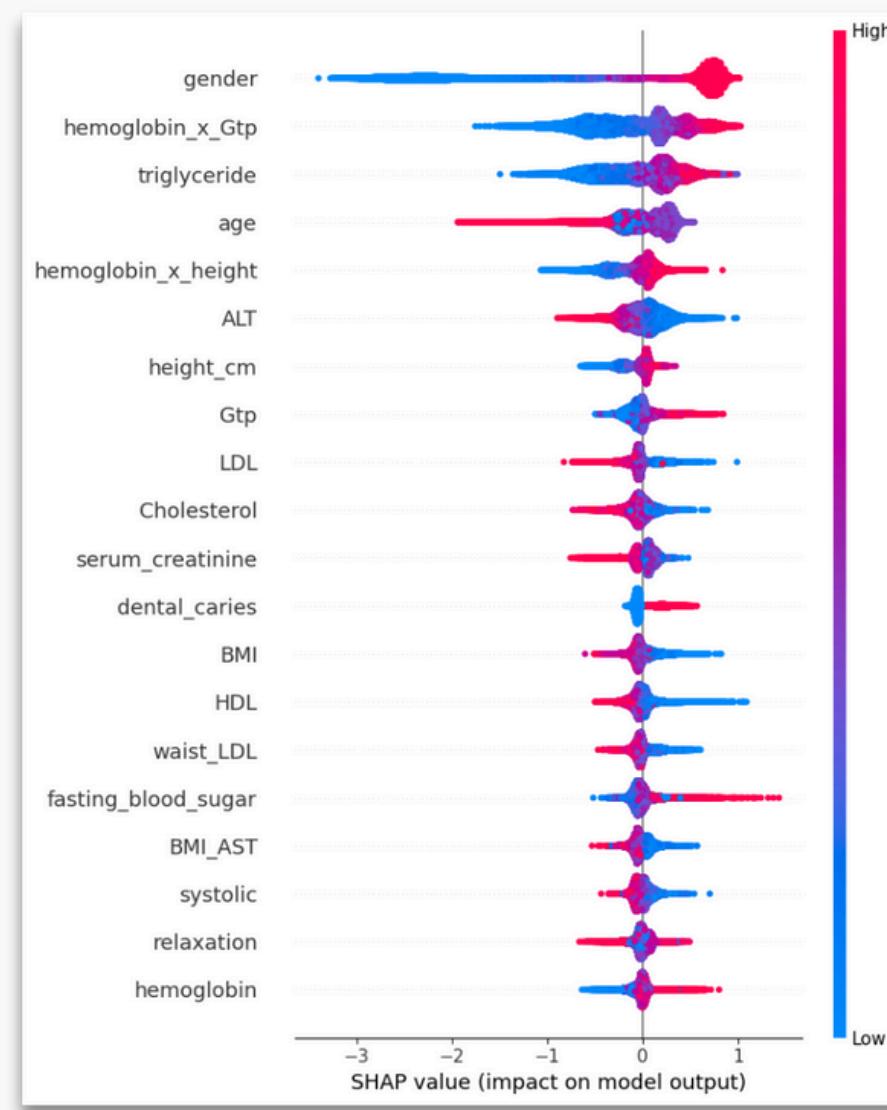
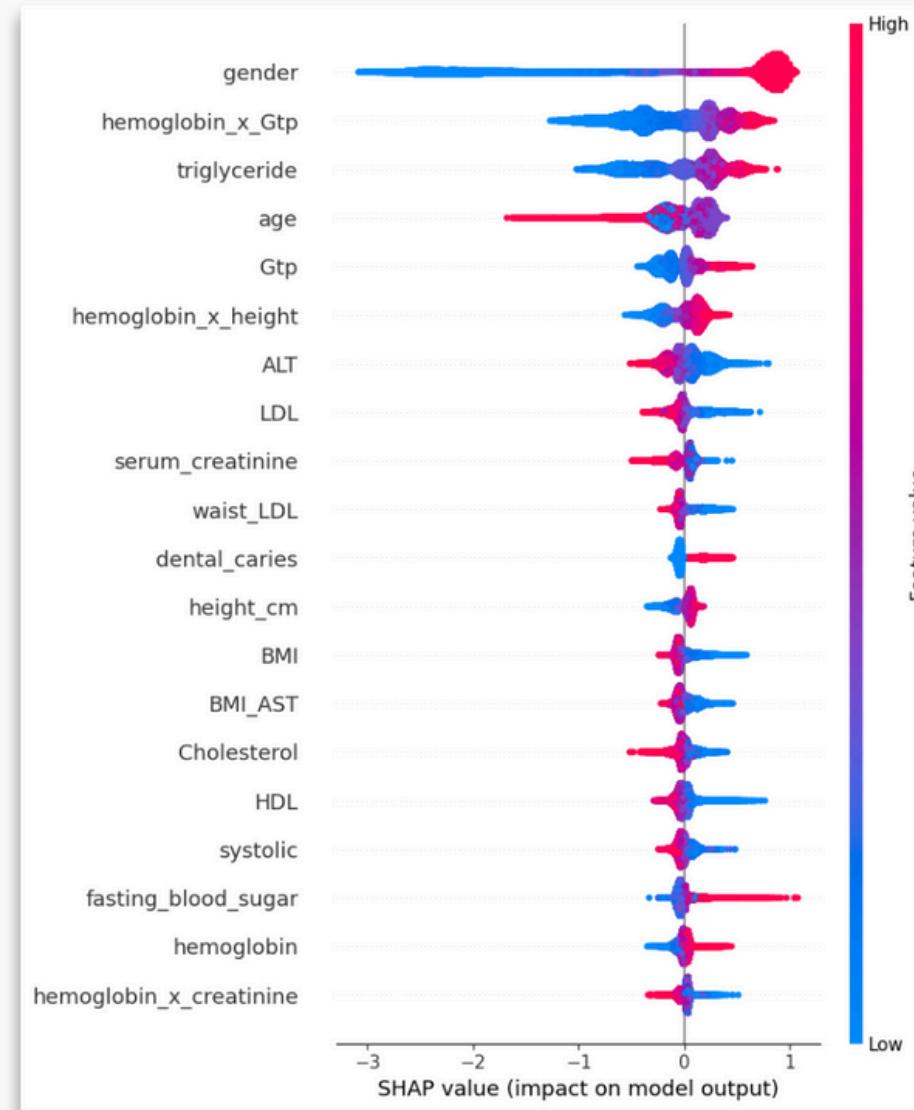
	feature	VIF
0	const	5434.680751
17	hemoglobin_x_weight	433.694009
20	weight_x_height	364.337228
16	hemoglobin_x_height	327.611693
15	hemoglobin_sq	298.860100
2	BMI	39.120038
14	waist_LDL	21.048839
11	LDL_risk	7.681881
3	WHTR	6.548142
8	BP_category	5.975081
5	MAP	5.841239
6	LDL_ratio	5.521221
1	gender	5.203417
7	BP_level	5.073143
9	total_cholesterol_risk	3.833736
19	hemoglobin_x_creatinine	2.723547
4	pulse_pressure	2.141147
13	BMI_AST	1.604848
10	HDL_risk	1.558739
18	hemoglobin_x_Gtp	1.525367
12	age_risk	1.372695

Model	Accuracy	Precision	Recall	F1	ROC-AUC
CatBoost	0.7811	0.7188	0.8205	0.7663	0.8641
LightGBM	0.7803	0.7190	0.8168	0.7648	0.8633
XGBoost	0.7784	0.7197	0.8081	0.7614	0.8624
RandomForest	0.7717	0.7112	0.8048	0.7551	0.8536
AdaBoost	0.7677	0.7020	0.8149	0.7542	0.8481
LogisticRegression	0.7441	0.7037	0.7167	0.7101	0.8299

# 6. 모델링과 성능 평가

## 6-3. FEATURE SELECTION (2/2) - 시각화

### SHAP를 활용한 중요 변수 식별



# 6. 모델링과 성능 평가

## 6-4. 양상을 모델 비교

모델 유형	구성 모델	선정 이유	기대 효과	ROC-AUC	성과 요약
<b>Soft Voting</b>	LGBM, XGBoost, CatBoost	서로 다른 구조의 모델 예측 확률 평균화로 보완적 특성 활용	정밀도-재현율 균형 확보 과적합 억제일반화 성능 향상	0.8688	가장 높은 AUC를 기록했으며, False Negative을 줄여 실제 예측 상황에서도 안정적으로 활용 가능할 것으로 예상
<b>Bagging</b>	단순형: DecisionTreeClassifier 배깅형: XGBoost (with sampling)	높은 분산의 단순 트리 모델 보완을 위해 배깅 적용. XGBoost에 배깅 효과를 부여하여 과적합 억제	분산 감소로 테스트 데이터에서도 성능 확보, 과적합 억제	0.8379 0.8672	XGBoost의 경우 예측 균형도 양호하며 높은 AUC를 기록했으나, 단일 모델 기반이므로 극한의 성능 향상에는 한계가 있을 것으로 예상
<b>Stacking</b>	LGBM, XGBoost, CatBoost + Logistic Regression	메타모델(LogReg)을 통한 비선형 결합 구조 학습	개별 모델의 약점을 메타모델이 보완하여 성능 향상	0.8684	Soft Voting과 유사한 수준의 성능. 복잡한 구조와 긴 학습 시간이 단점이라 비용-효과 면에서 트레이드오프 존재 고려
<b>Weighted Stacking (Optuna)</b>	LGBM, XGBoost, CatBoost	OOF(Out-of-Fold) 예측 기반 Optuna로 최적 가중치 탐색	정보 누수 방지 모델별 기여도 반영한 예측	0.8687	Soft Voting과 유사한 OOF AUC 기록했으나 실제 성능구조 대비에는 제한적

```
Soft Voting Classifier 평가 결과
Classification Report:
precision    recall    f1-score   support
          0       0.84      0.76      0.80     17921
          1       0.72      0.82      0.77     13931

accuracy                           0.78      31852
macro avg       0.78      0.78      0.78     31852
weighted avg    0.79      0.78      0.78     31852

ROC AUC Score: 0.8688
```

```
Bagging XGBoostClassifier 평가 결과
Classification Report:
precision    recall    f1-score   support
          0       0.84      0.75      0.80     17921
          1       0.72      0.82      0.77     13931

accuracy                           0.78      31852
macro avg       0.78      0.79      0.78     31852
weighted avg    0.79      0.78      0.78     31852

ROC AUC Score: 0.8672
```

```
Stacking 평가 결과
Classification Report:
precision    recall    f1-score   support
          0       0.83      0.78      0.80     17921
          1       0.73      0.79      0.76     13931

accuracy                           0.78      31852
macro avg       0.78      0.78      0.78     31852
weighted avg    0.79      0.78      0.78     31852

ROC AUC Score: 0.8684
```

```
[TEST] Weighted Stacking Optuna 평가 결과
Classification Report:
precision    recall    f1-score   support
          0       0.84      0.76      0.80     17921
          1       0.72      0.82      0.77     13931

accuracy                           0.78      31852
macro avg       0.78      0.79      0.78     31852
weighted avg    0.79      0.78      0.78     31852

ROC AUC Score: 0.8687
```

# 6. 모델링과 성능 평가

## 6-5. 최종 모델 선정

### 모델 선정 근거

#### SOFT VOTING 기반 양상별

- XGBOOST + LIGHTGBM + CATBOOST  
→ 스케일링 및 로그 변환 수행 X
- THRESHOLD 조정 없이 예측 확률값 그대로 사용
- ADVERSARIAL VALIDATION을 통해 데이터셋 간 분포 안정성 확인

1. 다양한 모델을 활용한 보완적 예측 구조에 기반한 안정적 성능 확보
2. 단순하면서도 모델 간 학습 시간과 구조적 복잡도를 최소화할 수 있는 설계
3. 실제 테스트셋에서 최고 수준의 AUC(0.8688)을 기록하며 다른 고비용 구조(STACKING, OPTUNA) 대비 성능 차이가 거의 없음

# 6. 모델링과 성능 평가

## 6-6. 이슈 트래킹

### 1. test 'id' 컬럼 순서 문제로 인한 성능 저하



### 2. 파생변수 생성 시 다중공선성 문제 발생

단계	상세
문제	로컬에선 AUC 0.86 이상, 캐글 제출 시 0.78~0.79로 급락
원인	예측값과 id를 수동 zip하는 과정에서 순서 불일치 발생
해결	sample_submission.csv 기준으로 id 유지 후 예측값 덮어쓰기
개선	캐글 점수 0.78 → 0.86+로 회복
인사이트	제출용 id 순서 보존 필수, 예측 전엔 제거 / 제출 전엔 복원

단계	상세
문제	new_features 리스트에 반복적으로 변수 추가 → 학습·테스트 데이터 분리 적용 시 동일 파생 변수 중복 생성 → VIF 확인 시 무한대(inf) 출력 → 다중공선성 경고
해결	리스트 자체는 유지하되, 중복 제거 방식으로 정리 (append 대신 set 또는 중복 제거 후 반환)

# 7. 프로젝트 인사이트

## 7-1. 시도 대비 결과



### 도메인 기반 파생 피처 생성

- gender 단일 피처 사용 시 가장 높은 AUC 달성
- 전체 피처 사용 시 성능 하락  
( $0.87382 \rightarrow 0.87194$ )  
=> 핵심 정보가 소수 변수에 집중되어 있음  
=> 생성한 다수 파생 피쳐는 모델 성능에 기여하지 않음



### 다중공선성 제거 기반 FEATURE SELECTION

- gender 단일 피처 사용 시보다 성능 낮음  
( $0.86308 \rightarrow 0.86222$ )
- 파생 피처 생성 후 VIF 기반 선택 수행  
=> 불필요한 피처 제거를 시도했으나  
유효 정보 부족으로 오히려 성능 최저치 기록



### 하이퍼파라미터 튜닝 및 모델 고도화

- 수동 설정 대비 자동 튜닝 모델 성능 저하
- 최적화된 threshold 적용 시 오히려 점수 하락.  
( $0.87355 \rightarrow 0.79159$ )  
=> 정보량이 적은 데이터에서는 과한 튜닝이 역효과  
=> 보수적인 수동 설정이 더 효과적인 전략으로 작용

# 7. 프로젝트 인사이트

## 7-2. 결과 분석

### 1. 데이터는 딥러닝 기반으로 합성된 인공 데이터

- 해당 대회의 데이터는 실제 환자 데이터를 수집한 것이 아닌 딥러닝 모델을 통해 생성된 합성 데이터
- 의미 있는 변수 간 상관관계가 인위적으로 설계되어 있거나, 실제 생리학적 인과 구조를 반영하지 않을 수 있음.
- 따라서 통계 기반 탐색 및 도메인 해석이 잘 작동하지 않음.

### 2. 주요 변수들의 정보량 부족

- age, height 컬럼은 정확한 나이가 아닌 5 단위로 구간화된 값  
→ 분류 결정에 필요한 미세한 경계 정보를 제공하지 못함.
- 시력, 청력, 체중, 간수치 등의 생체 신호는 흡연 여부와 직접적인 관성이 낮거나 매우 간접적인 지표
- gender만이 유의미했던 이유는 흡연 여부의 통계적 경향성이 실제로 성별에 따라 다르기 때문 (현실 반영)

### 3. 고차원 데이터지만 유효 피처는 거의 없음

- 데이터에 포함된 수십 개의 지표 중 실제로 label(smoking)과 통계적으로 의존적인 피처는 거의 없음.
- 다중공선성 제거/피처 선택을 수행해도 “잡음만 제거”되고 유효 신호는 여전히 부족함.

### 4. 모델 고도화가 적합하지 않았던 이유

- 파라미터 최적화는 유의미한 피쳐가 여러 개 존재할 때 유효함.
- 유의미한 변수가 성별 하나뿐이어서 모델이 학습할 정보가 제한적이었기 때문에 튜닝을 해도 generalization이 향상되지 않음.

# 7. 프로젝트 인사이트

## 7-3. 하이퍼파라미터 관련 차이점 분석

```
# 수동 설정 (AUC: 0.87382)
manual_params = {
    'colsample_bytree': 0.213,           # 소수 피처만 활용 → 노이즈 억제
    'subsample': 0.88,                  # 안정적인 샘플링
    'colsample_bylevel': 0.91,          # 전체 피처보다 부분 활용이 유리
    'min_child_samples': 230,           # 과도한 분할 방지 → 일반화 향상
    'min_child_weight': 21,             # 리프 분할 규제
    'n_estimators': 1000,               # 충분한 트리 수 → 수렴 유도
    'learning_rate': 0.019             # 낮은 학습률 → 안정적 학습
}

# 튜닝 결과 (AUC: 0.86507) -> 성능 하락
tuned_params = {
    'colsample_bytree': 0.6,            # 더 많은 변수 사용
    'subsample': 0.8,                 # 샘플링 규제가 약함
    'colsample_bylevel': 1.0,           # 전체 피처 활용
    'min_child_samples': 20,            # 분할 제한 약함
    'min_child_weight': 5,              # 규제 부족
    'n_estimators': 500,               # 트리 수 부족
    'learning_rate': 0.02              # 러닝레이트 유사
```



## 핵심 인사이트

- 이 대회 데이터는 고차원 + 저정보량 구조
- 많은 피처 중 실제 label과 관련된 정보는 매우 제한적
- 따라서, “**불필요한 변수 억제 + 분할 규제**” 전략이 성능 향상에 결정적이었음.

# 7. 프로젝트 인사이트

## 7-4. 결론 및 개선 방향

“모델링에 앞서, 데이터의 품질과 정보량을 높이는데 집중할 필요가 있다.”

- 01  모델보다는 데이터가 성능을 결정한다.
- 02  고차원 데이터라도 label을 설명할 수 있는 핵심 정보가 없으면 모델링은 무의미하다.
- 03  합성 데이터의 경우 특히 피처의 의미나 도메인 기반 접근이 잘 작동하지 않을 수 있다.



# 7. 프로젝트 인사이트

## 7-5. 분석 과정에서 얻은 인사이트

01



### End-to-End

- 「데이터 이해 → 전처리 → 피처 엔지니어링 → 모델 학습 → 성능 평가 → 성능 향상 방안」
- 데이터 이해부터 성능 향상까지 전 과정을 직접 수행하여, 문제 해결 프로세스를 단계별로 경험하고 적용할 수 있었음.

02



### 팀 협업과 효율적 소통

- 코랩, 구글 드라이브, 깃허브, 노션, 캔바 등 다양한 협업 도구를 활용하여, 팀원 간 원활한 커뮤니케이션과 공동 산출물도 출이 가능하였음.

03



### 데이터 특성에 대한 심화 이해

- 데이터 분석을 통해 의미 있는 파생 변수를 생성하여, 피처별 성능 기여도 분석의 중요성을 직접 체감하였음.
- 예측력 향상에는 도메인 기반 해석이 핵심적임을 확인할 수 있었음.

04



### EDA에서 시각화의 중요성

- 수치만으로는 놓치기 쉬운 이상치와 변수 간 분포 차이를 히트맵, 바이올린플롯, 페어플롯 등의 시각화 기법을 통해 직관적으로 파악하여, 데이터를 더욱 정밀하게 분석할 수 있었음.

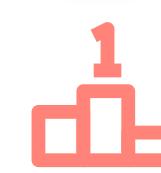
05



### Test 파일 검토의 중요성

- 데이터 전처리 과정에서 테스트 파일의 ID와 행 순서가 변형될 수 있음을 경험하였고, 모델 성능에 예상치 못한 영향을 줄 수 있어 테스트와 제출 파일 검토가 반드시 필요함을 깨달았음.

06



### 다양한 모델 비교를 통한 전략 수립

- LightGBM, XGBoost 등 다양한 모델을 실험하고 성능을 비교하여, 모델 특성과 결과 해석을 바탕으로 앙상블 전략의 당위성과 효과를 확인하여 적용할 수 있었음.



감사합니다.