



# 분류 보고서 양식

## 머신 러닝을 이용한 예측 분석

### 1. 프로젝트 개요

#### 1-1. 주제

생체 신호를 이용한 흡연자 상태의 이진 예측 (23.11.14 대회 마감)

Binary Prediction of Smoker Status using Bio-Signals

Playground Series - Season 3, Episode 24

[k https://www.kaggle.com/competitions/playground-series-s3e24/overview](https://www.kaggle.com/competitions/playground-series-s3e24/overview)

- 평가 지표 : **ROC Curve 아래 면적 (AUC)**
- 달성 목표: 상위 10%(191위)

#### 1-2. 주제 선정의 배경

- 흡연은 건강에 해로운 습관으로 인식되며 흡연으로 인한 질병과 사망률이 상당히 높다고 알려져 있어 이에 대한 대책 필요
- 흡연은 고혈압, 공복 혈당 수치의 변화, 콜레스테롤 수치의 이상, 혈액 내 헤모글로빈 농도의 증가 등과 관련이 있을 수 있으며 이러한 건강 지표들은 흡연이 건강에 미치는 영향을 평가하는 데 중요한 지표로 활용 가능
- 효과적인 공중보건 정책의 수립과 건강 프로그램의 개발에 활용 가능

#### 1-3. 본 프로젝트의 활용 방안 제시

- 생체 신호를 분석하여 흡연자의 특정 상태를 인식하여 금연 프로그램 지원

- 흡연량이나 흡연 빈도와 같은 정보를 기반으로 개인화된 건강 조언을 제공하는 시스템 구축

## 2. 프로젝트 수행 절차 및 방법

### 2-1. 데이터 설명

- 총 22개의 피쳐 변수와 1개의 타겟 변수로 구성

	feature	데이터 타입	결측값	고유향	max	min
0	age	int64	0	18	85.0	20.0
1	height(cm)	int64	0	15	190.0	130.0
2	weight(kg)	int64	0	29	135.0	30.0
3	waist(cm)	float64	0	548	129.0	51.0
4	eyesight(left)	float64	0	20	9.9	0.1
5	eyesight(right)	float64	0	18	9.9	0.1
6	hearing(left)	int64	0	2	2.0	1.0
7	hearing(right)	int64	0	2	2.0	1.0
8	systolic	int64	0	128	233.0	71.0
9	relaxation	int64	0	94	146.0	40.0
10	fasting blood sugar	int64	0	259	423.0	46.0
11	Cholesterol	int64	0	279	445.0	55.0
12	triglyceride	int64	0	393	999.0	8.0
13	HDL	int64	0	123	359.0	4.0
14	LDL	int64	0	286	1860.0	1.0
15	hemoglobin	float64	0	144	21.1	4.9
16	Urine protein	int64	0	6	6.0	1.0
17	serum creatinine	float64	0	34	11.6	0.1
18	AST	int64	0	196	1090.0	6.0
19	ALT	int64	0	230	2914.0	1.0
20	Gtp	int64	0	444	999.0	2.0
21	dental caries	int64	0	2	1.0	0.0
22	smoking	int64	0	2	1.0	0.0

• 피쳐 데이터

Feature		세부설명
age	나이	
height	신장	
weight	체중	
waist	허리	
eyesight(left)	시력(왼쪽)	
eyesight(right)	시력(오른쪽)	
hearing(left)	청각(왼쪽)	
hearing(right)	청각(오른쪽)	
systolic	혈압	심장이 수축할 때 동맥이 받는 압력의 양
relaxation	기분전환 / 혈압	심장이 이완할 때 동맥이 받는 압력의 양
fasting blood sugar	공복 혈당	<ul style="list-style-type: none"> <li>• 식사 전 혈액 내 포도당 농도를 나타냄</li> <li>• 고혈당은 당뇨병의 초기 증상 중 하나이며, 심각한 건강 문제를 초래할 수 있음</li> </ul>
cholesterol	콜레스테롤	<ul style="list-style-type: none"> <li>• 혈액 속의 지질류 중 하나</li> <li>• 동맥경화는 혈관 벽에 지방이 쌓여 혈관이 좁아지는 현상으로, 고 콜레스테롤 수치는 이를 촉진할 수 있음</li> </ul>
triglyceride	트리글리세리드	<ul style="list-style-type: none"> <li>• 지방의 주요 형태 중 하나</li> <li>• 과다한 트리글리세라이드는 심혈관 질환의 위험을 높일 수 있음</li> </ul>
HDL	고밀도 지단백질	HDL은 "좋은" 콜레스테롤로 알려져 있으며, 높은 HDL 수치는 심혈관 질환의 위험을 낮출 수 있음
LDL	저밀도 지단백질	<ul style="list-style-type: none"> <li>• 혈관 벽에 콜레스테롤이 쌓여 동맥경화의 원인이 되는 '나쁜 콜레스테롤'</li> <li>• 높은 수치가 지속되면 심근경색, 뇌경색 등의 질병이 발생할 위험이 증가</li> <li>• LDL 수치가 높은 사람이 흡연하면 그 위험도는 더욱 높아져 주의</li> </ul>
hemoglobin	헤모글로빈	<ul style="list-style-type: none"> <li>• 적혈구의 일부</li> <li>• 비흡연자의 혈중 헤모글로빈 수치는 1% 정도인데 반해 흡연자의 경우 5% 이상</li> </ul>
urine protein	소변 단백질	<ul style="list-style-type: none"> <li>• 소변 내에 과도한 단백질이 섞여 나오는 것</li> <li>• 단백뇨는 당뇨, 고혈압과 같은 만성 질환에서 나타날 수 있음</li> </ul>

Feature		세부설명
		<ul style="list-style-type: none"> <li>• 단백뇨 수치의 증가는 신장 손상 정도가 증가함을 의미</li> <li>• 흡연이 단백뇨 위험도를 증가 시키는 것과 관련이 있음</li> </ul>
serum creatinine	혈청 크레아티닌	<ul style="list-style-type: none"> <li>• 콩팥에 의해 변하지 않고 배설되는 근육 대사의 부산물</li> <li>• 콩팥 기능에 대해 가장 일반적으로 사용되는 지표</li> </ul>
AST	글루타민산	<ul style="list-style-type: none"> <li>• 신체가 아미노산을 분해하는데 도움이 되는 효소</li> <li>• 수치가 증가하면 간에 이상을 의미</li> </ul>
ALT	글루타민산	단백질 효소의 한 종류, 단백질을 간 세포의 에너지로 전환시키는데 도움을 줌
GTP	구아노신 삼인산	<ul style="list-style-type: none"> <li>• 혈액 내 효소</li> <li>• 간이나 담관 손상시 수치가 증가</li> </ul>
dental caries	충치 여부	충치 여부

- 타겟 데이터

Feature		세부설명
smoking	흡연 여부	0: 비흡연 / 1: 흡연

### 3. 데이터 전처리

#### 3-1. 데이터 전처리 계획

- 모델링을 위해 고유값을 가진 ID 피쳐 제거
- 결측값 확인
- 중복 데이터 확인
- 이상치 확인
- 의미있는 피쳐 변수 생성

#### 3-2. 데이터

##### ▼ Data

## Binary Prediction of Smoker Status using Bio-Signals

[Binary Prediction of Smoker Status using Bio-Signals.zip](#)

[smoking.csv](#)

```
train = pd.read_csv("/content/train.csv")
train_dataset = pd.read_csv("/content/train_dataset.csv")
test = pd.read_csv("/content/test.csv")
submission = pd.read_csv("/content/sample_submission.csv")
```

```
train.shape = (159256, 23)
train_dataset.shape = (38984, 23)
test.shape = (106171, 22)
submission.shape = (106171, 2)
```

- 전처리 후 train / test data
  - train.shape = (192723, 23)

	feature	데이터 타입	결측값	고유향	max	min
0	age	int64	0	18	85.0	20.0
1	height(cm)	int64	0	15	190.0	130.0
2	weight(kg)	int64	0	29	135.0	30.0
3	waist(cm)	float64	0	548	129.0	51.0
4	eyesight(left)	float64	0	20	9.9	0.1
5	eyesight(right)	float64	0	18	9.9	0.1
6	hearing(left)	int64	0	2	2.0	1.0
7	hearing(right)	int64	0	2	2.0	1.0
8	systolic	int64	0	128	233.0	71.0
9	relaxation	int64	0	94	146.0	40.0
10	fasting blood sugar	int64	0	259	423.0	46.0
11	Cholesterol	int64	0	279	445.0	55.0
12	triglyceride	int64	0	393	999.0	8.0
13	HDL	int64	0	120	150.0	4.0
14	LDL	int64	0	193	200.0	1.0
15	hemoglobin	float64	0	144	21.1	4.9
16	Urine protein	int64	0	6	6.0	1.0
17	serum creatinine	float64	0	25	3.0	0.1
18	AST	int64	0	95	100.0	6.0
19	ALT	int64	0	150	150.0	1.0
20	Gtp	int64	0	290	300.0	2.0
21	dental caries	int64	0	2	1.0	0.0
22	smoking	int64	0	2	1.0	0.0

- test.shape = (106171, 22)

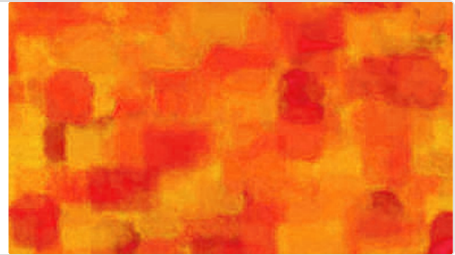
	feature	데이터 타입	결측값	고유통값	max	min
0	age	int64	0	18	85.0	20.0
1	height(cm)	int64	0	16	190.0	135.0
2	weight(kg)	int64	0	26	130.0	30.0
3	waist(cm)	float64	0	508	127.7	51.0
4	eyesight(left)	float64	0	20	9.9	0.1
5	eyesight(right)	float64	0	18	9.9	0.1
6	hearing(left)	int64	0	2	2.0	1.0
7	hearing(right)	int64	0	2	2.0	1.0
8	systolic	int64	0	114	213.0	71.0
9	relaxation	int64	0	78	140.0	40.0
10	fasting blood sugar	int64	0	224	423.0	46.0
11	Cholesterol	int64	0	227	369.0	66.0
12	triglyceride	int64	0	392	548.0	8.0
13	HDL	int64	0	106	148.0	18.0
14	LDL	int64	0	182	200.0	1.0
15	hemoglobin	float64	0	132	21.1	5.0
16	Urine protein	int64	0	6	6.0	1.0
17	serum creatinine	float64	0	23	3.0	0.1
18	AST	int64	0	93	100.0	6.0
19	ALT	int64	0	142	150.0	1.0
20	Gtp	int64	0	271	300.0	2.0
21	dental caries	int64	0	2	1.0	0.0

### 3-3. 데이터 수집 및 전처리

- 추가 data 확보

## Smoker Status Prediction using Bio-Signals

<https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals>



- 효율적인 모델링을 위해 고유값을 가진 ID 피처를 index로 사용

```
1 train = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/train.csv", index_col="id")
2 train_dataset = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/train_dataset.csv")
3 test = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/test.csv", index_col="id")
4 submission = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/sample_submission.csv")
```

- 기존 train data와 추가로 수집한 data 병합

```
1 train = pd.concat([train, train_dataset])
2 train
```

	age	height(cm)	weight(kg)	waist(cm)	eyesight(left)	eyesight(right)	hearing(left)	hearing(right)	systolic	relaxation	...	HDL	LDL	hemoglobin	Urine protein	serum creatinine	AST
0	55	165	60	81.0	0.5	0.6	1	1	135	87	...	40	75	16.5	1	1.0	22
1	70	165	65	89.0	0.6	0.7	2	2	146	83	...	57	126	16.2	1	1.1	27
2	20	170	75	81.0	0.4	0.5	1	1	118	75	...	45	93	17.4	1	0.8	27
3	35	180	95	105.0	1.5	1.2	1	1	131	88	...	38	102	15.9	1	1.0	20
4	30	165	60	80.5	1.5	1.0	1	1	121	76	...	44	93	15.4	1	0.8	19
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
38979	40	165	60	80.0	0.4	0.6	1	1	107	60	...	61	72	12.3	1	0.5	18
38980	45	155	55	75.0	1.5	1.2	1	1	126	72	...	76	131	12.5	2	0.6	23
38981	40	170	105	124.0	0.6	0.5	1	1	141	85	...	48	138	17.1	1	0.8	24
38982	40	160	55	75.0	1.5	1.5	1	1	95	69	...	79	116	12.0	1	0.6	24
38983	55	175	60	81.1	1.0	1.0	1	1	114	66	...	64	137	13.9	1	1.0	18

198240 rows x 23 columns



### - 전처리 전

train.shape =

(159256, 23)

original.shape = (38984, 23)

-

### 전처리 후

train.shape = (198240, 9)

- 중복 데이터 제거



```
1 train_df.shape  
(198240, 23)  
  
1 train_df.drop_duplicates().shape  
(192723, 23)
```



- 총 5517개 제거

- 결측값 확인

```

1 train_df.isnull().sum()
age                0
height(cm)         0
weight(kg)         0
waist(cm)          0
eyesight(left)     0
eyesight(right)    0
hearing(left)      0
hearing(right)     0
systolic           0
relaxation         0
fasting blood sugar 0
Cholesterol        0
triglyceride       0
HDL                0
LDL                0
hemoglobin         0
Urine protein      0
serum creatinine   0
AST                0
ALT                0
Gtp                0
dental caries      0
smoking            0
dtype: int64

1 test.isnull().sum()
age                0
height(cm)         0
weight(kg)         0
waist(cm)          0
eyesight(left)     0
eyesight(right)    0
hearing(left)      0
hearing(right)     0
systolic           0
relaxation         0
fasting blood sugar 0
Cholesterol        0
triglyceride       0
HDL                0
LDL                0
hemoglobin         0
Urine protein      0
serum creatinine   0
AST                0
ALT                0
Gtp                0
dental caries      0
dtype: int64

```

- GTP, HDL, LDL, ALT, AST, Serum creatinine 해당 변수에 대한 Outlier 제거

```

1 # train 데이터에 대한 임계값 설정
2 train_df['Gtp'] = train_df['Gtp'].clip(lower = 0, upper = 300)
3 train_df['HDL'] = train_df['HDL'].clip(lower = 0, upper = 150)
4 train_df['LDL'] = train_df['LDL'].clip(lower = 0, upper = 200)
5 train_df['ALT'] = train_df['ALT'].clip(lower = 0, upper = 150)
6 train_df['AST'] = train_df['AST'].clip(lower = 0, upper = 100)
7 train_df['serum creatinine'] = train_df['serum creatinine'].clip(lower = 0, upper = 3)
8
9 # test 데이터에 대한 임계값 설정
10 test['Gtp'] = test['Gtp'].clip(lower = 0, upper = 300)
11 test['HDL'] = test['HDL'].clip(lower = 0, upper = 150)
12 test['LDL'] = test['LDL'].clip(lower = 0, upper = 200)
13 test['ALT'] = test['ALT'].clip(lower = 0, upper = 150)
14 test['AST'] = test['AST'].clip(lower = 0, upper = 100)
15 test['serum creatinine'] = test['serum creatinine'].clip(lower = 0, upper = 3)

```



- discussion에서 많이 사용 되는 outlier 제거 방법
- 각 피처에 해당하는 정상 범위를 바탕으로 min / max를 clip함수로 outlier를 대체

### 3-4. 활용 라이브러리 등 기술적 요소

- ✓ pandas → 모듈
- ✓ numpy
- ✓ matplotlib
- ✓ seaborn
- ✓ warnings
- ✓ random
- ✓ copy
- ✓ lightgbm
- ✓ catboost
- ✓ xgboost
- ✓ sklearn
- ✓ optuna

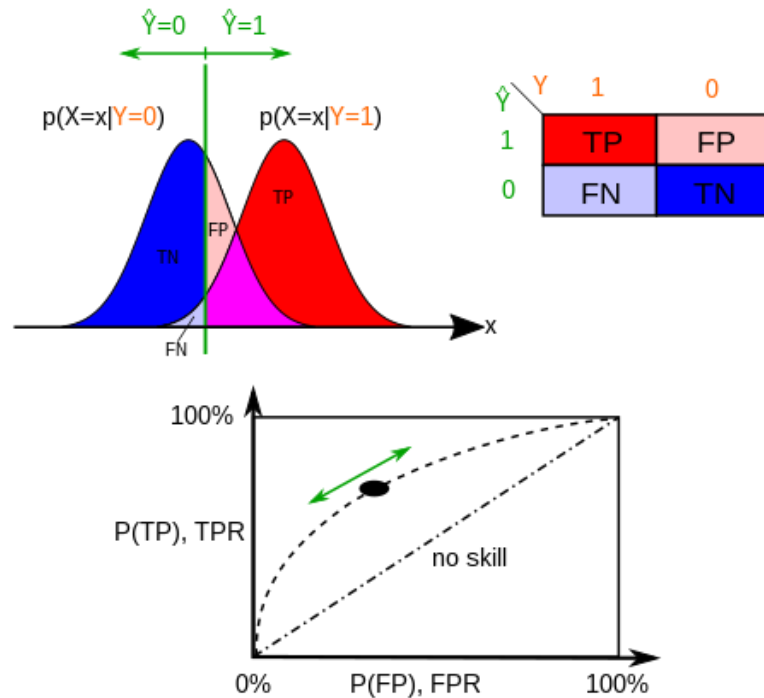
### 3-5. 프로젝트에서 분석한 내용

- ✓ 결측치 확인
- ✓ 중복값 확인
- ✓ 데이터 타입 확인
- ✓ 이상치 확인
- ✓ 병합된 데이터 shape 확인
- ✓ 병합된 데이터 타입 확인
- ✓ 임계값 조정 전과 후 데이터들의 분포도 시각화
- ✓ 병합 데이터의 흡연여부 시각화
- ✓ 각 데이터들 상관관계 시각화
- ✓ 각 데이터들 분포도 시각화
- ✓ 각 데이터별 특성 분포도 시각화
- ✓ ROC 곡선 생성
- ✓ 피처 중요도 시각화

## 4. 기초 평가

### 4-1. 지표평가

- ROC-AUC



## ◦ ROC-Curve

▼ FPR(False Positive Rate)의 변화에 따른 TPR(True Positive Rate)의 변화를 나타내는 곡선

- FPR(False Positive Rate)
  - 실제 Negative(음성, 0)를 잘못 예측한 비율
  - $FP / (FP + TN)$
- TPR(True Positive Rate)
  - 실제 Positive(양성, 1)가 정확히 예측되어야 하는 수준
  - 재현율(Recall), 민감도(Sensitivity)라고도 불린다.
  - $TP / (FN + TP)$

- 이진 분류 모델의 성능을 시각화 하는 도구
- 왼쪽 위 모서리에 가까울수록 모델의 성능이 우수하다고 판단

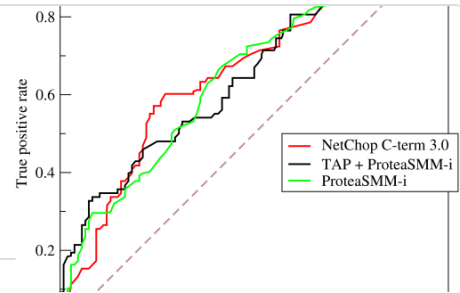
## ◦ AUC (Area Under Curve)

- ROC Curve의 아래 면적을 나타내는 지표이며 모델의 전반적인 성능을 요약
- 0과 1 사이의 값을 가지며 1에 가까울수록 모델의 성능이 우수하다고 판단

### Receiver operating characteristic

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier model at varying threshold values.

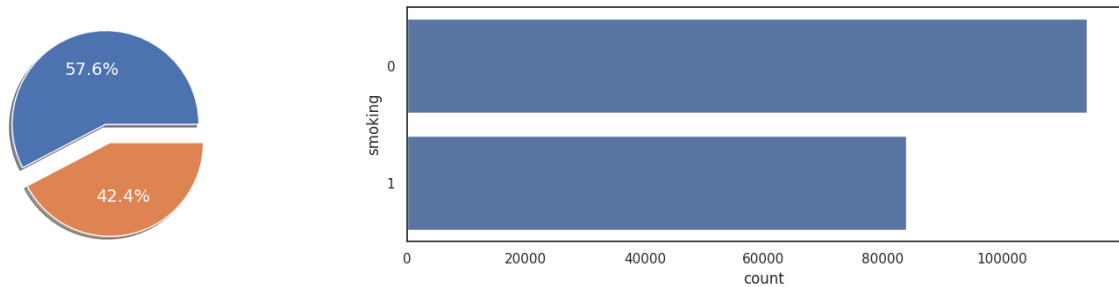
W [https://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic)



## 4-2 . 시각화- 추이, 편차, 구성비율

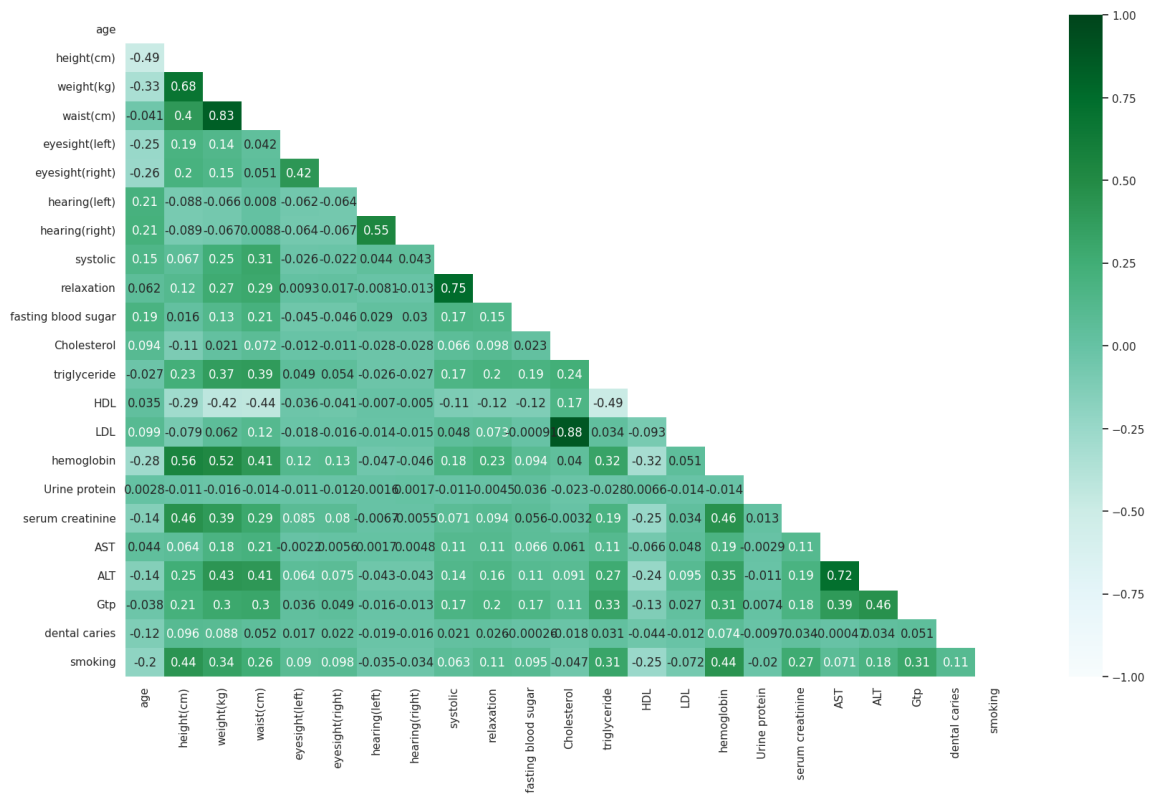
### ▼ Target(smoking) 분포 확인

Smoking Distribution in Train



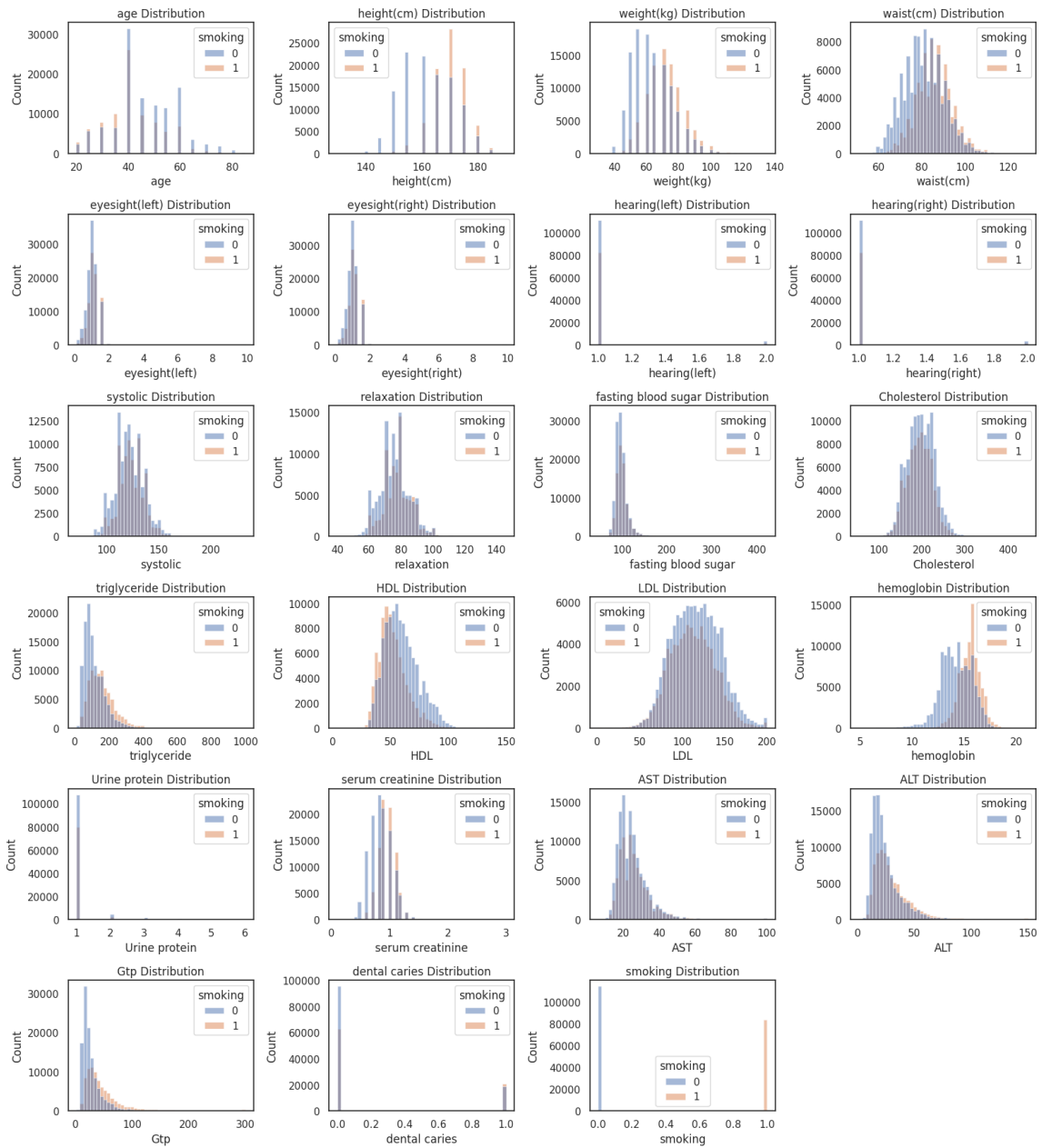
- 비흡연자 57.6% / 흡연자 42.4%
- 균형적인 타겟의 분포

### ▼ 변수 간의 상관 관계



- 각 피쳐들 간 상관 관계가 높은지 낮은지 확인
- 키 / 몸무게 / 허리둘레 피쳐가 유의미한 상관관계 확인
- 혈압과 콜레스테롤 피쳐가 유의미한 상관관계 확인
- 간 수치에 관련한 피쳐들이 서로 유의미한 상관관계 확인

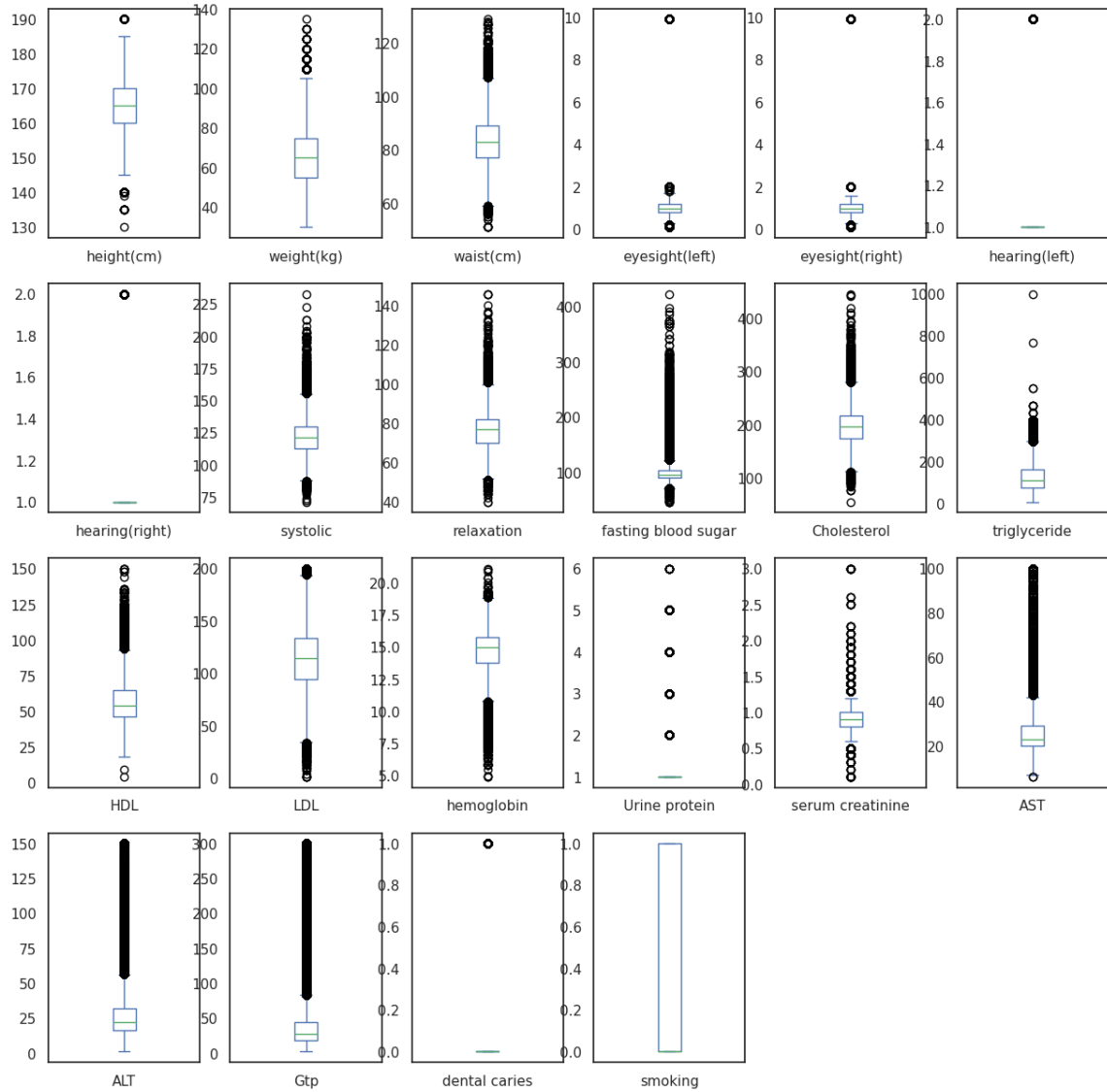
## ▼ 각 데이터 분포



- 혈액 관련 피쳐들이 정규분포의 형태를 띄고 있는 것을 확인

## ▼ 각 데이터의 이상치



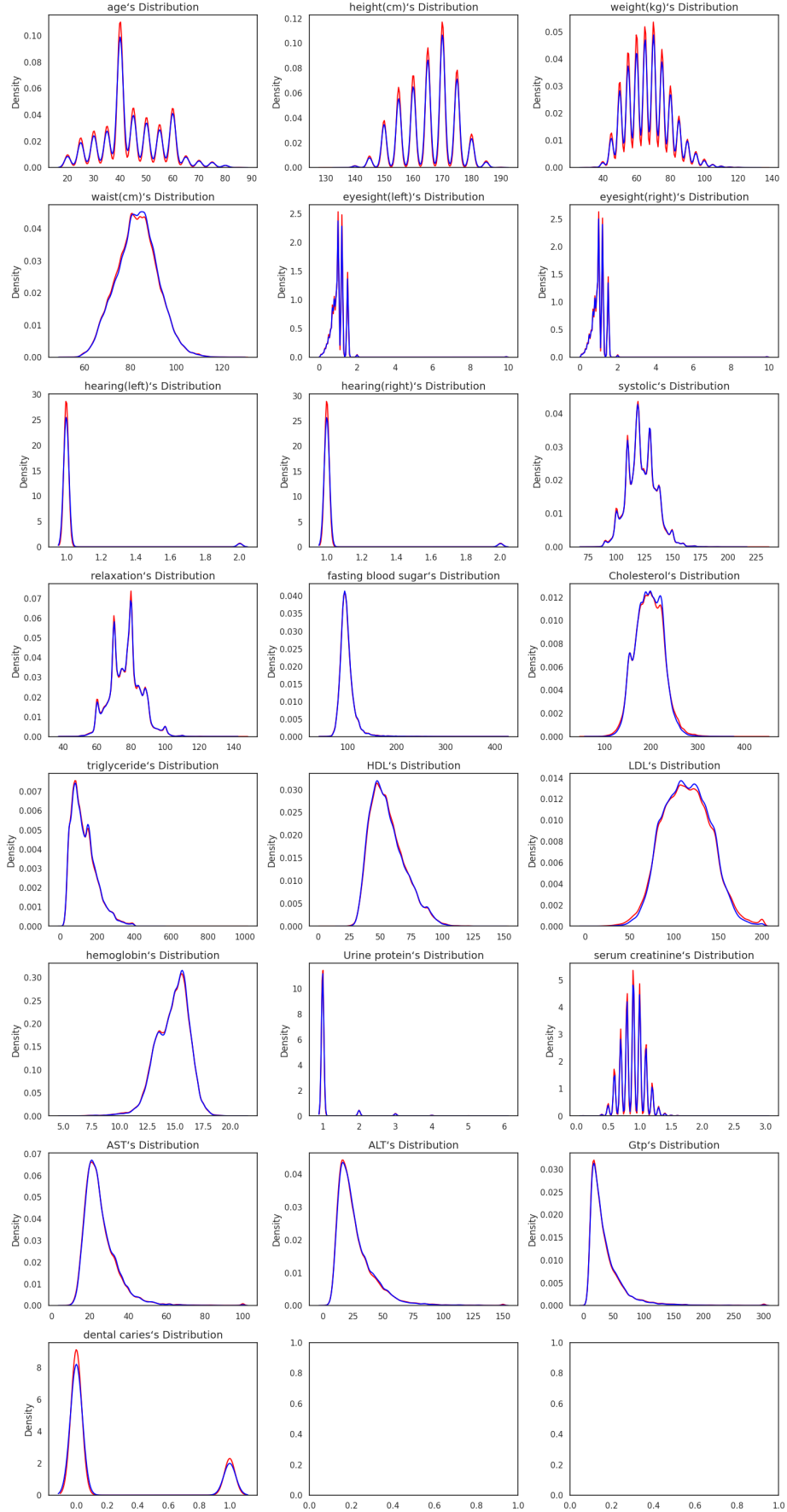


- IQR 기준 이상치 확인
- 도메인적 지식을 바탕으로 이상치 제거 여부 보류

## ▼ 각 데이터 세트 분포

## Distribution of Feature per Dataset

— Train  
— Test

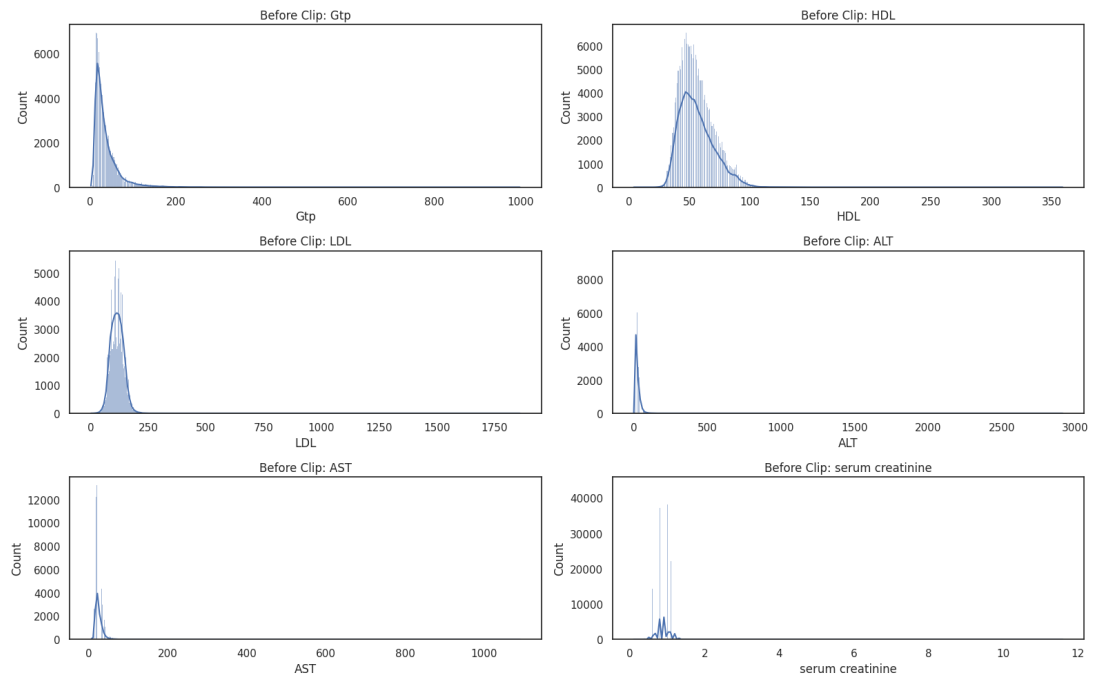




- train 데이터는 빨간색 / test 데이터는 파란색
- train / test 모두 비슷한 분포

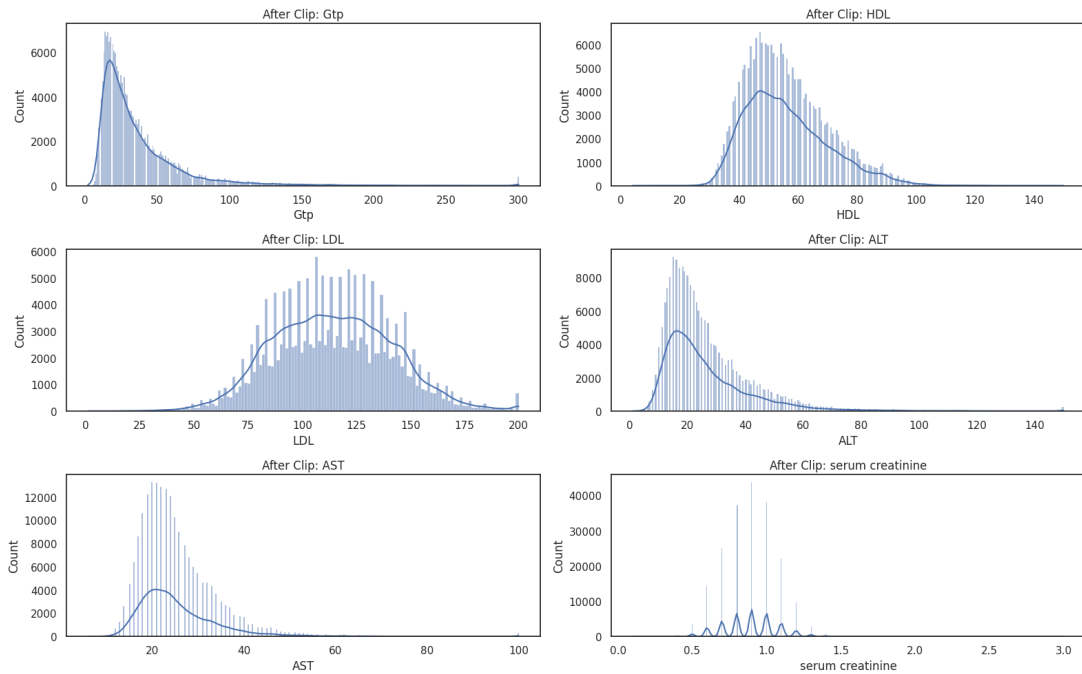
▼ 특정 피처에 대한 임계값 조정 전 / 후 분포

• 조정 전



- 오른쪽으로 긴 꼬리를 형성
- 대체적인 값들이 정규분포를 따르는 모습을 보이고 있으나 이상치를 가진 값들이 존재

• 조정 후



- 이상치를 임계값으로 대체

## 5. 피쳐 엔지니어링

### 5-1 피쳐 엔지니어링

#### 1. clip()함수를 사용하여 임계값 조정

```
1 # train 데이터에 대한
2 train['Gtp'] = train['Gtp'].clip(lower = 0, upper = 300)
3 train['HDL'] = train['HDL'].clip(lower = 0, upper = 150)
4 train['LDL'] = train['LDL'].clip(lower = 0, upper = 200)
5 train['ALT'] = train['ALT'].clip(lower = 0, upper = 150)
6 train['AST'] = train['AST'].clip(lower = 0, upper = 100)
7 train['serum creatinine'] = train['serum creatinine'].clip(lower = 0, upper = 3)
8
9 # test 데이터에 대한
10 test['Gtp'] = test['Gtp'].clip(lower = 0, upper = 300)
11 test['HDL'] = test['HDL'].clip(lower = 0, upper = 150)
12 test['LDL'] = test['LDL'].clip(lower = 0, upper = 200)
13 test['ALT'] = test['ALT'].clip(lower = 0, upper = 150)
14 test['AST'] = test['AST'].clip(lower = 0, upper = 100)
15 test['serum creatinine'] = test['serum creatinine'].clip(lower = 0, upper = 3)
```

- train, test데이터의 피쳐['Gtp', 'HDL', 'LDL', 'ALT', 'AST', 'serum creatinine']들의 값이 한 쪽으로 몰려있는 이상치 확인 → clip()함수 사용하여 임계값 조정




















## 2. ID피처를 인덱스로 설정

```
1 train = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/train.csv", index_col="id")
2 train_dataset = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/train_dataset.csv")
3 test = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/test.csv", index_col="id")
4 submission = pd.read_csv("/content/drive/MyDrive/테킷/9.파이널프로젝트-12/data/sample_submission.csv")
```

index\_col="id" → 데이터 프레임의 행(row)을 식별하는 고유한 식별자(identifier)로 사용

## 6. 모델 학습

### 6-1. 프로젝트에 사용했던 방법들

-  (0.86803) submission\_5Fold\_depth(12)
-  (0.86907) submission\_stacker\_robust
-  (0.87041) submission\_stacker\_standard
-  (0.87050) submission\_stacker\_minmax
-  (0.87194) submission\_5Fold(2)(임계값 -50)
-  (0.87386) submission\_5Fold(X)
-  (0.87413) submission\_5Fold(3)(임계값 +50)
-  (0.87413) submission\_5Fold
-  (0.87474) submission\_LGB
-  (0.87553) submission\_LGB\_핫코딩\_minmax(rate=0.3)
-  (0.87626) submission\_LGB(X)
-  (0.87760) submission\_LGB\_핫코딩\_minmax(rate=0.12)
-  (0.87767) submission\_LGB\_핫코딩
-  (0.87770) submission\_LGB\_핫코딩\_minmax(boosting=gdbt)
-  (0.87784) submission\_LGB\_fold\_원핫인코딩
-  (0.87795) lgb\_submission\_minmax
-  (0.87797) lgb\_submission\_minmax
-  (0.87802) lgb\_submission\_minmax
-  (0.87805) lgb\_submission\_final\_8

- 앙상블, 다양한 모델, 다양한 스케일링, 다양한 Fold 수, 하이퍼 파라미터 값 수정, 임계값 조정 등 여러가지 방법으로 학습

### <결론>

- LGBM모델이 확실히 좋은 성능의 모델인 것을 확인

- 교차 검증에서 Fold 횟수는 7, random\_state 횟수는 42가 가장 적합
- LGBM을 단일 모델로 사용하여 각각의 fold에서 훈련된 모델들을 사용하여 테스트 데이터에 대한 예측을 생성

## 7. 머신 러닝 결과

### 7-1. 결과 분석



lgb\_submission\_final\_8.csv

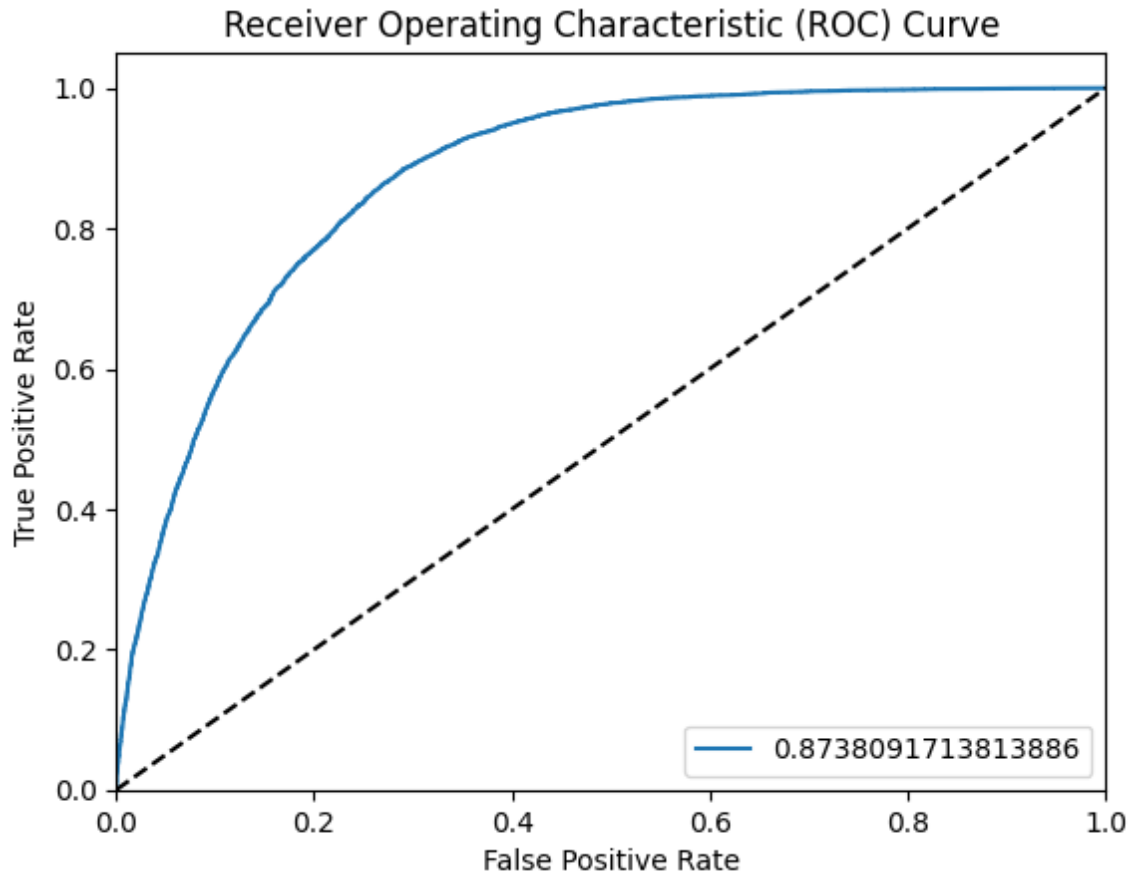
완료(마감일 이후) · 6일 전

0.87805

187	▼ 27	산		0.87806	5	5개월
188	▲ 5	완웨이뉴에		0.87806	5	6개월
189	▲ 5	에티옌A		0.87806	삼	5개월
190	▲ 5	슈밤 차반		0.87806	5	5개월
191	▼ 11	낸시		0.87799	11	6개월
192	▲ 19	파이살 알스레이드		0.87798	삼	5개월
193	▲ 7	은행 계좌		0.87797	10	6개월
194	▲ 7	유티아오		0.87794	24	5개월

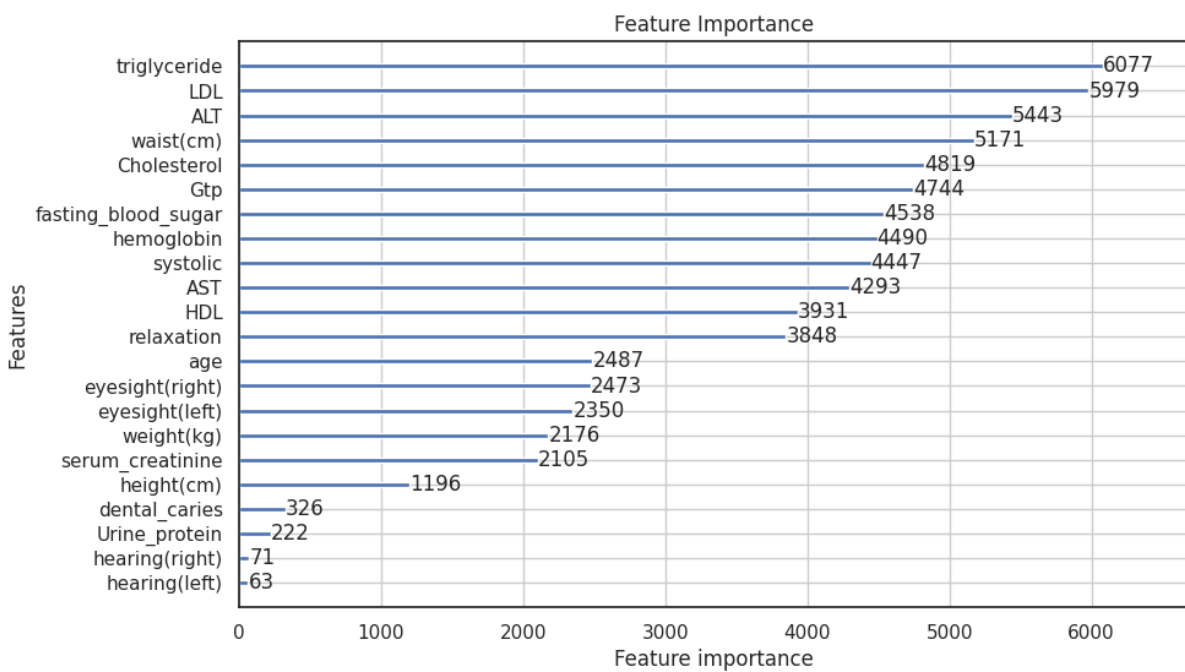
- 0.87805로 190등과 191등 사이이므로 상위 10%(192등)안에 진입

### 7-2. ROC Curve



- ROC 곡선 그래프 확인 결과 0.87로 좋은 모델로 판단

### 7-3. 피쳐 중요도 결과



- 흡연과 관련이 높은 혈액 / 혈관과 관련이 있는 피쳐들이 중요하게 나왔으며 그 뒤로 시력, 무게, 청력이 뒤따르는 것을 확인

## 8. 프로젝트 회고 및 개선점

### 8-1. 피드백

제출 전에는 이곳이 공백입니다.

발표 후 QnA 시간에 나온 질문과 피드백을 모두 작성해 주세요.

듣는 즉시 바로 작성하면 빠뜨리지 않고 모두 적을 수 있을 거예요! 이때 개선점으로 넘어가도 좋을 반영할 부분을 발견했다면 최종 제출 전에 그 부분 위주로 정리하는 것도 좋아요. 그리고 발표 시간에 적극적으로 질문과 피드백을 주고 받으면 서로의 성장에 무척 도움이 되겠죠?

### 8-2. 회고

### 8-3. 개선점

### 8-4. 추후 개선 계획

개선점에 대한 회고 이후, 가능하다면 실제 액션 계획도 세워보세요. 포트폴리오에서 '개선 시도/경험'은 아주 긍정적인 요소로 작용한답니다.

## 8. 부록

### 8-1. 참고자료

<분석>

<https://github.com/Koda98/smoker-status-prediction/tree/main>

<https://www.kaggle.com/code/xxxxxyyy80008/smoker-status-prediction-lightgbm-baseline-no-fe>



<https://www.kaggle.com/code/arunklenin/ps3e24-smoking-cessation-prediction-binary>

## 8-2. 출처

<분석>

<https://www.kaggle.com/competitions/playground-series-s3e24/overview>

<https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals>