

🚬 Smoking & Health - 로지스틱 회귀 분석 (혼동행렬, ROC, AUC 포함)

문제

다음은 나이, BMI, 흡연 여부에 따른 심장병 발생 여부 데이터이다. 로지스틱 회귀를 통해 변수의 영향력을 분석하고, 예측 성능 평가를 수행하시오.

변수 설명:

- Age: 나이
- BMI: 체질량지수
- Smoker: 흡연 여부 (0: 비흡연, 1: 흡연)
- Heart_Disease: 심장병 발생 여부 (0: 없음, 1: 있음)

문제:

1. 로지스틱 회귀 모델을 구성하여 Heart_Disease 를 예측하시오.
2. 각 변수의 오즈비를 구하고 해석하시오.
3. 혼동행렬을 작성하고 정확도, 정밀도, 재현율을 해석하시오.
4. ROC 곡선 및 AUC 값을 계산하여 모델의 분류 성능을 평가하시오.

✓ 정답 코드

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, classification_report, roc_auc_score,
roc_curve
import matplotlib.pyplot as plt
```

```

df = pd.read_csv("smoking_health_expanded.csv")
X = df[['Age', 'BMI', 'Smoker']]
y = df['Heart_Disease']

```

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

```
X_train = sm.add_constant(X_train)
```

```
X_test = sm.add_constant(X_test)
```

~~#~~ model = sm.Logit(y_train, X_train).fit()

pred_prob = model.predict(X_test) → 예측 확률

pred = (pred_prob > 0.5).astype(int) → 이진 분류 예측

오즈비

```
print(np.exp(model.params))
```

혼동행렬

```
cm = confusion_matrix(y_test, pred)
```

```
print(cm)
```

분류 리포트

```
print(classification_report(y_test, pred))
```

ROC, AUC

```
fpr, tpr, thresholds = roc_curve(y_test, pred_prob)
```

```
auc_score = roc_auc_score(y_test, pred_prob)
```

```
print("AUC:", auc_score)
```

Pred_Prob	Pred
0. 83	1
0. 12	0
0. 52	1

이전 분류기 결과와 바꾸어 봄.
0.3-14-0.1 등...

💡 해설

- Smoker 변수의 오즈비가 1 보다 크다면, 흡연자는 심장병 발생 확률이 높은 것으로 해석할 수 있다.
- 혼동행렬의 TP, TN, FP, FN 을 기준으로 정확도(Accuracy), 정밀도(Precision), 재현율(Recall)을 해석해야 한다.
- ROC 곡선은 분류 임계값 변화에 따른 민감도(TPR)와 위양성률(FPR)을 시각화한 그래프이며, AUC 값은 모델의 전체적인 분류 성능을 수치화한 것이다.
- 이 모델의 AUC 는 약 0.5 로, 1 에 가까울수록 좋은 분류 성능을 의미한다.

혼동행렬 및 성능 평가

혼동행렬:

```
[[TN: 8  FP: 2]  
 [FN: 1  TP: 1]]
```

분류 리포트:

	precision	recall	f1-score	support
0	0.89	0.80	0.84	10
1	0.33	0.50	0.40	2
accuracy		0.75	0.75	12
macro avg	0.61	0.65	0.62	12
weighted avg	0.80	0.75	0.77	12

ROC & AUC

AUC 값: 0.5

ROC 곡선은 TPR(F1)과 FPR의 관계를 시각화한 곡선으로, AUC 값이 0.5에 가까우면 랜덤 분류 수준, 1에 가까우면 완벽한 분류자이다.