로지스틱 회귀 분석을 활용한 심장 질환 주요 변수 예측

의료데이터분석 및 활용 실무

정보융합학부 2020204051 김서경

CONTENT

01. 프로젝트 개요 데이터 선정 배경

심장 질환 국내 사망원인 2위

(단위: 인구 10만 명당 명)

순위	사망원인	사망률	'22년 순위 대비
1	악성신생물(암)	166.7	-
2	심장 질환	64.8	-
3	폐렴	57.5	⬆(+1)
4	뇌혈관 질환	47.3	⬆(+1)
5	고의적 자해(자살)	27.3	⬆(+1)
6	알츠하이머병	21.7	⬆(+1)
7	당뇨병	21.6	⬆(+1)
8	고혈압성 질환	15.6	⬆(+1)
9	패혈증	15.3	⬆(+2)
10	코로나19	14.6	⬇(-7)

젊은 층 심장 질환 환자 가파른 증가세

뉴스홈 | 최신기사

심장질환자 4년새 20% 늘었다…20대는 무려 33% 급증

송고시간 | 2023-11-07 17:02

성서호 기자

진료비도 40% 급증해 지난해 2조5천억원 달해

심평원, 최근 5년간 심장질환 진료 분석…환자 수 183만여명

01. 프로젝트 개요 데이터 선정 배경

심장 질환 국내 사망원인 2위

(단위: 인구 10만 명당 명)

순위	사망원인	사망률	'22년 순위 대비
1	악성신생물(암)	166.7	-
2	심장 질환	64.8	
3	폐렴		
4	뇌혈관 질환		
5	고의적 자해		
6	알츠하이머병		
7	당뇨병		
8	고혈압성 질환	15.6	⬆(+1)
9	패혈증	15.3	⬆(+2)
10	코로나19	14.6	⬇(-7)

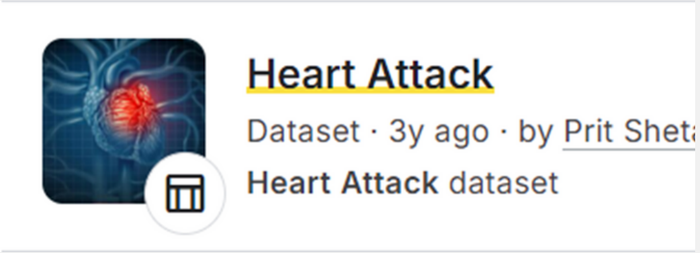
젊은 층 심장 질환 환자 가파른 증가세

로지스틱 회귀 분석을 통해
심장 질환 발병의 주요 원인을 분석하자!

02. 데이터 EDA 데이터 개요

kaggle ‘Health care: Heart attack possibility’

https://www.kaggle.com/datasets/pritsheta/heart-attack



02. 데이터 EDA 데이터 개요

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

주요 변수 설명

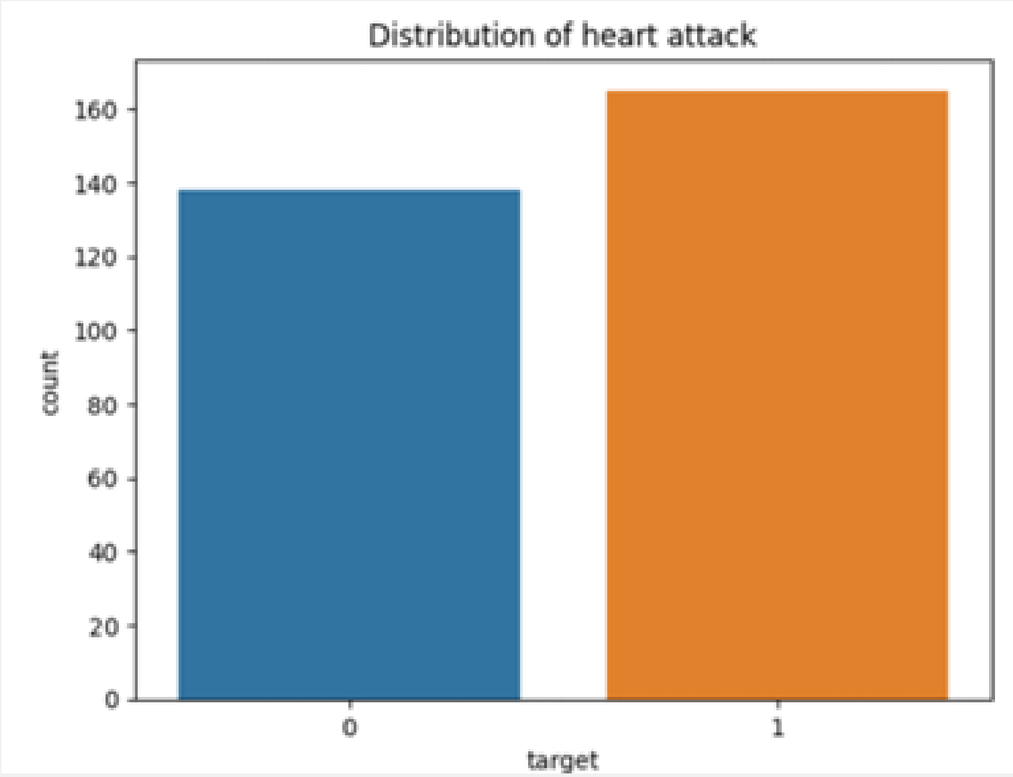
- Age (나이) : 환자의 나이를 나타내는 연속형 데이터
- Sex (성별) : 환자의 성별을 나타내며, 1은 남성, 2는 여성
- Cp (흉통 유형) : 환자가 경험한 흉통 유형을 네 가지로 구분
  0 : 전형적인 협심증
  1 : 비전형적인 협심증
  2 : 비협심증성 통증
  3 : 무증상
- Trestbps (휴식 시 혈압) : 병원 입원 시 측정된 환자의 혈압 (단위: mmHg)
- Chol (콜레스테롤) : 혈액 내 총 콜레스테롤 수치 (단위: mg/dL)
- Fbs (공복 혈당) : 공복 시 혈당이 120mg/dL를 초과하는지 여부, 1(초과함) / 0(정상)
- Restecg (휴식 시 심전도 결과) : 휴식 상태에서 측정된 심전도 결과
  0 : 정상
  1 : ST-T파 이상
  2 : 좌심실 비대 가능성
- Thalach (최대 심박수) : 극한 운동 시 도달할 수 있는 최대 심박수 (단위: bpm)
- Exang (운동 유발 협심증) : 운동 중 협심증 발생 여부, 1(있음) / 0(없음)
- Oldpeak (운동으로 유발된 ST 분절 변화) : 운동과 휴식 간 ST 분절의 차이 (단위: mm)
- Slope (ST 분절 기울기) : 운동 중 ST 분절의 기울기를 나타냄
  0 : 상승형
  1 : 평형형
  2 : 하강형
- Ca (주요 혈관 수) : 혈관 조영술로 관찰된 주요 혈관의 수 (범위: 0~4)
- Thal (핵의학 스캔 결과) : 심장 상태를 나타냄, 1(질환있음) / 2(가역적 결함)
- **Target (심장 질환 여부) : 심장 질환의 유무를 나타냄, 0(질환 없음) / 1(질환 있음)**

02. 데이터 EDA 데이터 개요

```
df.isnull().sum()
```

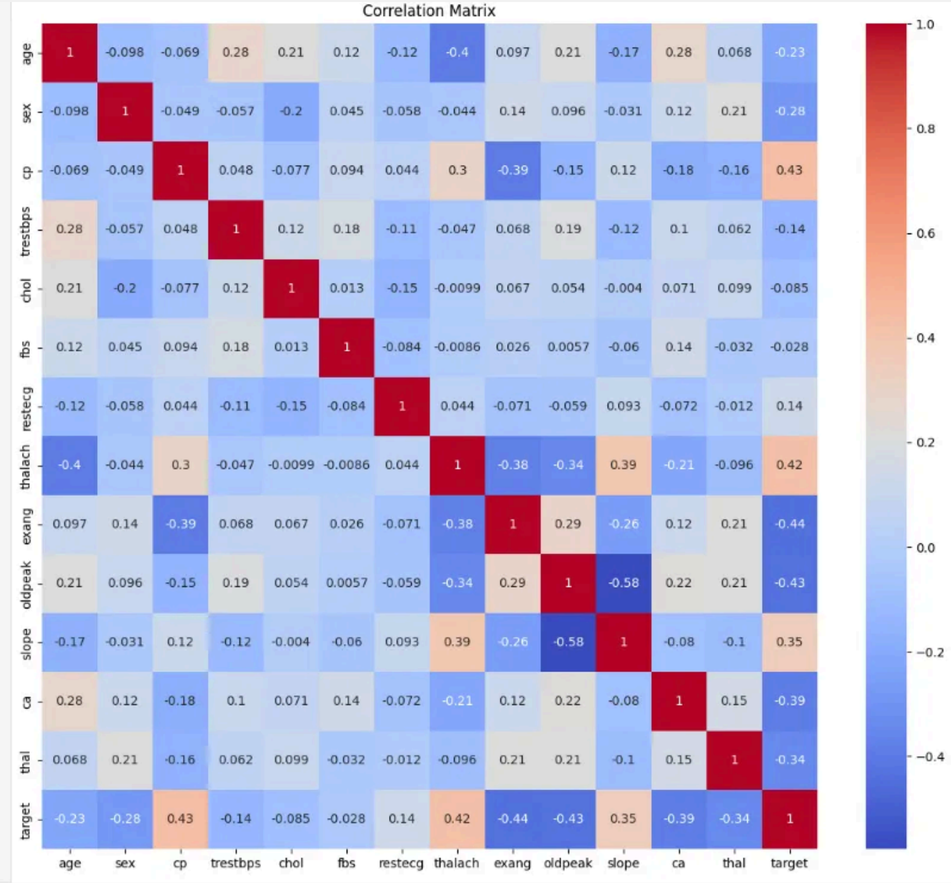
```
[28]:

age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

14개 변수, 결측치 없음



종속 변수 분포 – balanced

02. 데이터 EDA 데이터 개요



변수 간 상관계수 행렬

'target' 과의 주요 관계

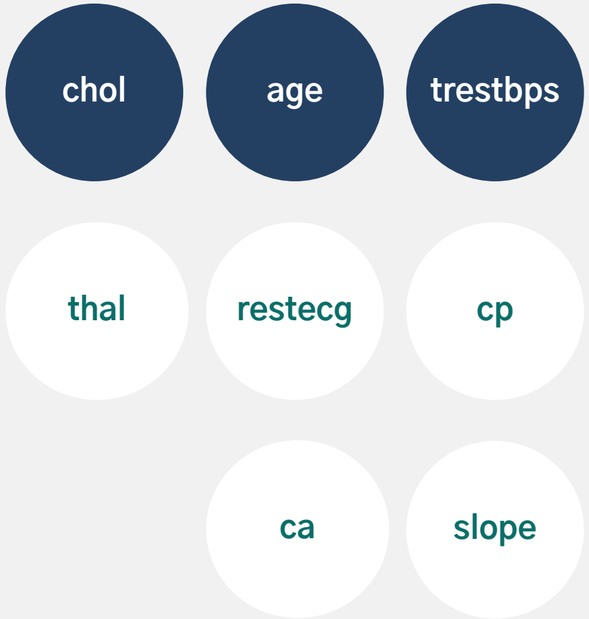
양의 상관 관계

- cp(가슴통증) 0.43
- thalach(최대 심박수) 0.42
- slope(ST segment기울기) 0.35

음의 상관관계

- exang(운동으로 유발된 협심증) –0.44
- oldpeak(ST 분절기울기) –0.43
- ca(주요 혈관 수) –0.39

03. ANOVA 분석 연속형 변수 범주화



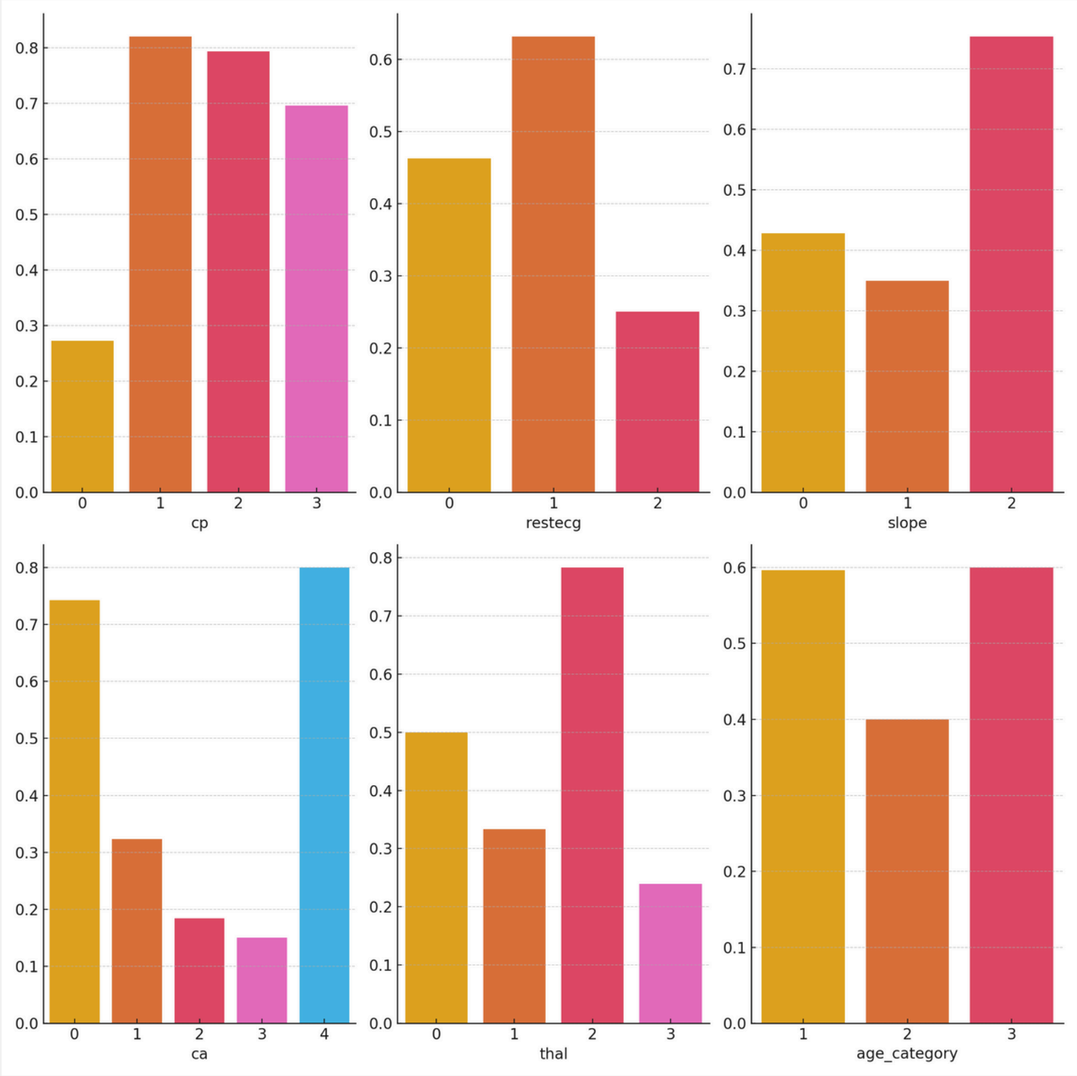
범주	1	2	3	4
age	60세 미만	60-69세	70-79세	
trestbps	120 미만	120-139	140-159	160이상
chol	200 미만	200-239	240 이상	

질병관리청. (n.d.). 일반건강정보. Retrieved December 9, 2024, from https://health.kdca.go.kr/healthinfo/biz/health/gnrlzHealthInfo/gnrlzHealthInfo/gnrlzHealthInfoView.do?cntnts_sn=6054)

질병관리청. (n.d.). 일반건강정보. Retrieved December 9, 2024, from https://health.kdca.go.kr/healthinfo/biz/health/gnrlzHealthInfo/gnrlzHealthInfo/gnrlzHealthInfoView.do?cntnts_sn=5300)

03. ANOVA 분석 분석 결과

변수	F값	p-값	유의미 여부 (p < 0.05)
cp	36.7	< 0.001	유의미
restecg	4.73	0.043	유의미
slope	28.0	< 0.001	유의미
ca	24.7	< 0.001	유의미
thal	30.7	0.002	유의미
age_group	4.55	0.021	유의미
trestbps_group	2.28	0.084	유의미하지 않음
chol_group	1.88	0.157	유의미하지 않음



범주 별 심장 질환 발생 비율

04. 로지스틱 회귀 분석 I pre

8:2 RANDOM SAMPLING

X, y

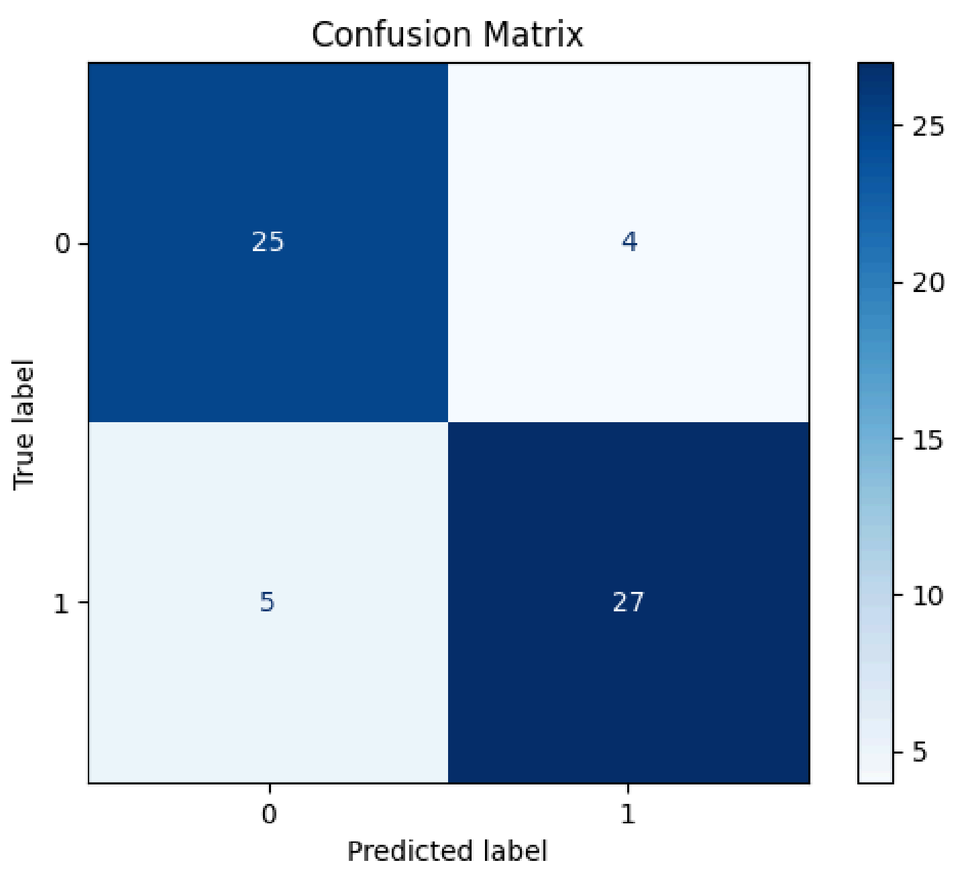
```
X = df.drop("target", axis=1)
y = df["target"]
```

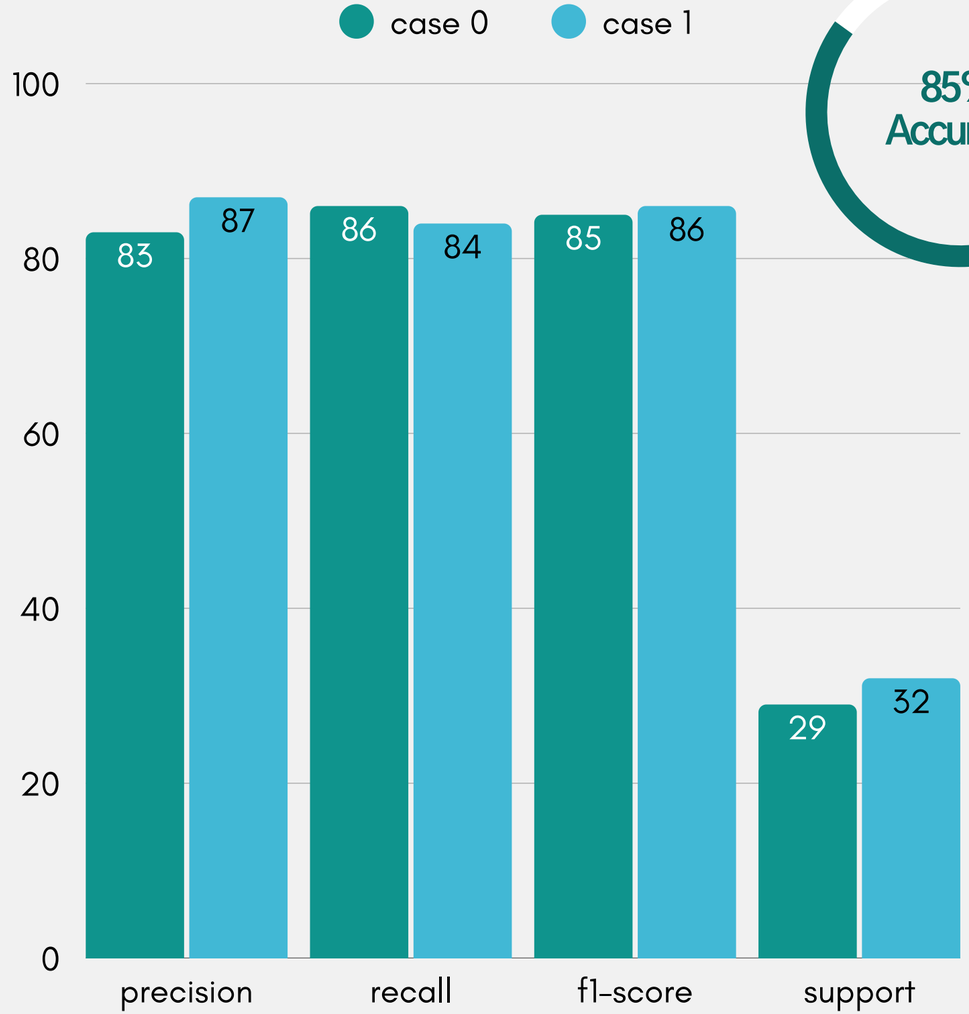
StandardScaler()

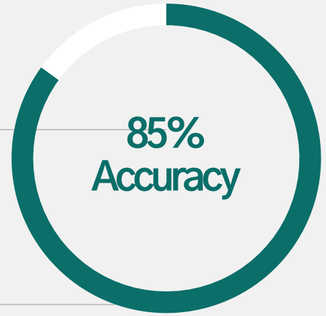
로지스틱 회귀는 가중치(weight)에 민감

-> StandardScaler를 사용하여 데이터를 표준화

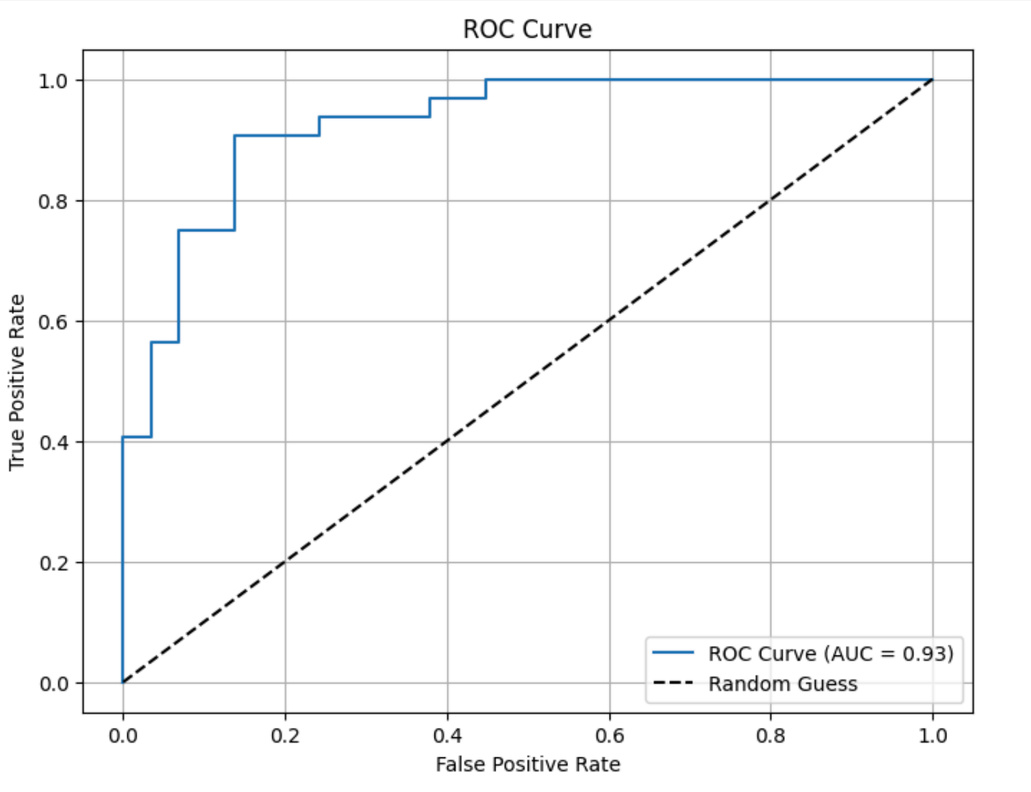
04. 로지스틱 회귀 분석 I classification_report(y_test, y_pred)







04. 로지스틱 회귀 분석 I ROC Curve



X

잘못된 양성 예측 비율 (False Positive Rate, FPR)

Y

올바른 양성 예측 비율, (True Positive Rate, TPR), 즉 재현율(Recall)

ROC Curve

분류 임계값(threshold)을 조정하면서 FPR과 TPR의 관계를 나타낸 곡선

-> 왼쪽 위로 치우쳐져 있어, FPR이 낮은 상태에서도 높은 TPR을 유지

-> 93%의 확률로 양성과 음성을 올바르게 구분할 수 있음 (AUC = 0.93)

04. 로지스틱 회귀 분석 I Cross-Validation (5-Fold)

```
교차 검증 정확도: [0.90163934 0.85245902 0.7704918  0.81666667 0.85      ]
평균 정확도: 83.83%
```

모델의 일반화 가능성을 평가

StratifiedKFold를 사용해 클래스 비율을 유지한 상태로 데이터를 나누고 데이터를 5개의 Fold로 나누어 5번 학습 및 검증을 시행

-> 성능 평균 83.83%로 일반화 가능성 높음

-> 77%~90 % 편차 있음

=> VIF 확인

05. Feature Selection Variance Inflation Factor(VIF)

	Variable	VIF
0	const	212.998773
1	age	1.443937
2	sex	1.231356
3	cp	1.397152
4	trestbps	1.180747
5	chol	1.152971
6	fbs	1.087698
7	restecg	1.066721
8	thalach	1.653567
9	exang	1.440147
10	oldpeak	1.744666
11	slope	1.662325
12	ca	1.290729
13	thal	1.191528
14	target	2.072754

다중공선성이 있는 경우, 특정 변수의 계수가 왜곡되거나 작아질 수 있음

VIF > 10

다중공선성이 확인됨

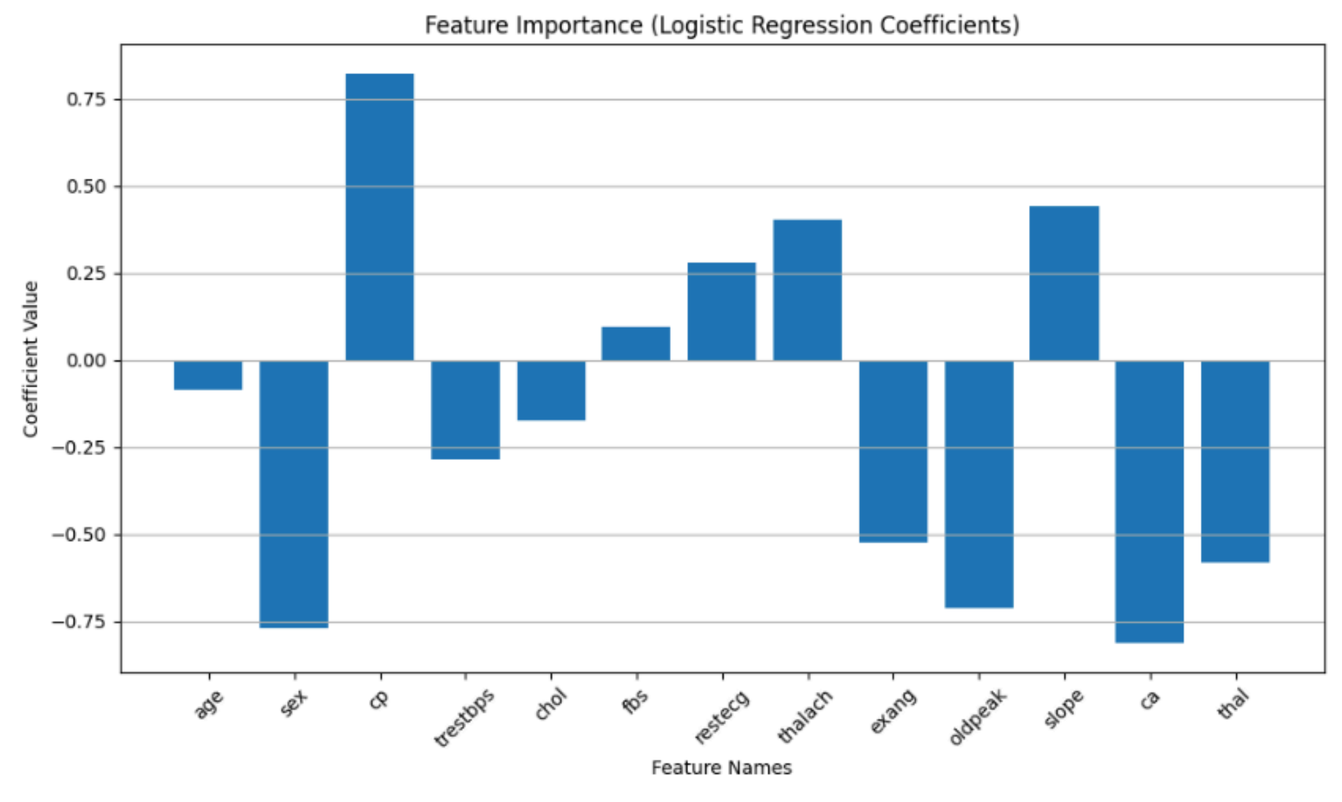
모두 1에 가까움

다중공선성 거의 없는 데이터

=> 컬럼 선택 불가

=> Logistic Regression Coefficients 확인

05. Feature Selection Logistic Regression Coefficients



모델의 주요 변수 식별

계수 > 0

해당 변수 증가 -> 종속 변수 예측 확률 증가

계수 < 0

해당 변수 증가 -> 종속 변수 예측 확률 감소

|계수| 클수록 변수 중요도 큼

=> coef 의 통계적 유의성 확인

05. Feature Selection sm.Logit(y_train, X_train_const).fit()

```
Optimization terminated successfully.
         Current function value: 0.348242
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                  242
Model:                          Logit   Df Residuals:                      228
Method:                           MLE   Df Model:                           13
Date:                Sun, 08 Dec 2024   Pseudo R-squ.:                  0.4940
Time:                        17:28:09   Log-Likelihood:                -84.274
converged:                       True   LL-Null:                       -166.55
Covariance Type:            nonrobust   LLR p-value:                 2.359e-28
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0330      0.206     -0.160      0.873      -0.437       0.371
age           -0.0795      0.228     -0.349      0.727      -0.527       0.368
sex           -0.8286      0.243     -3.411      0.001      -1.305      -0.352
cp             0.8862      0.219      4.048      0.000       0.457       1.315
trestbps      -0.3091      0.205     -1.509      0.131      -0.710       0.092
chol          -0.1951      0.211     -0.924      0.356      -0.609       0.219
fbs            0.1047      0.227      0.461      0.645      -0.340       0.550
restecg        0.3106      0.208      1.493      0.135      -0.097       0.718
thalach        0.4277      0.277      1.542      0.123      -0.116       0.971
exang         -0.5434      0.216     -2.513      0.012      -0.967      -0.120
oldpeak       -0.7639      0.297     -2.576      0.010      -1.345      -0.183
slope          0.4700      0.243      1.931      0.053      -0.007       0.947
ca            -0.8753      0.232     -3.773      0.000      -1.330      -0.421
thal          -0.6156      0.210     -2.928      0.003      -1.028      -0.204
==============================================================================
```

통계적으로 유의하지 않은 계수를 가진 변수가 다수(P>0.05)

=> RandomForestClassifier

05. Feature Selection RandomForestClassifier

Feature Importances:

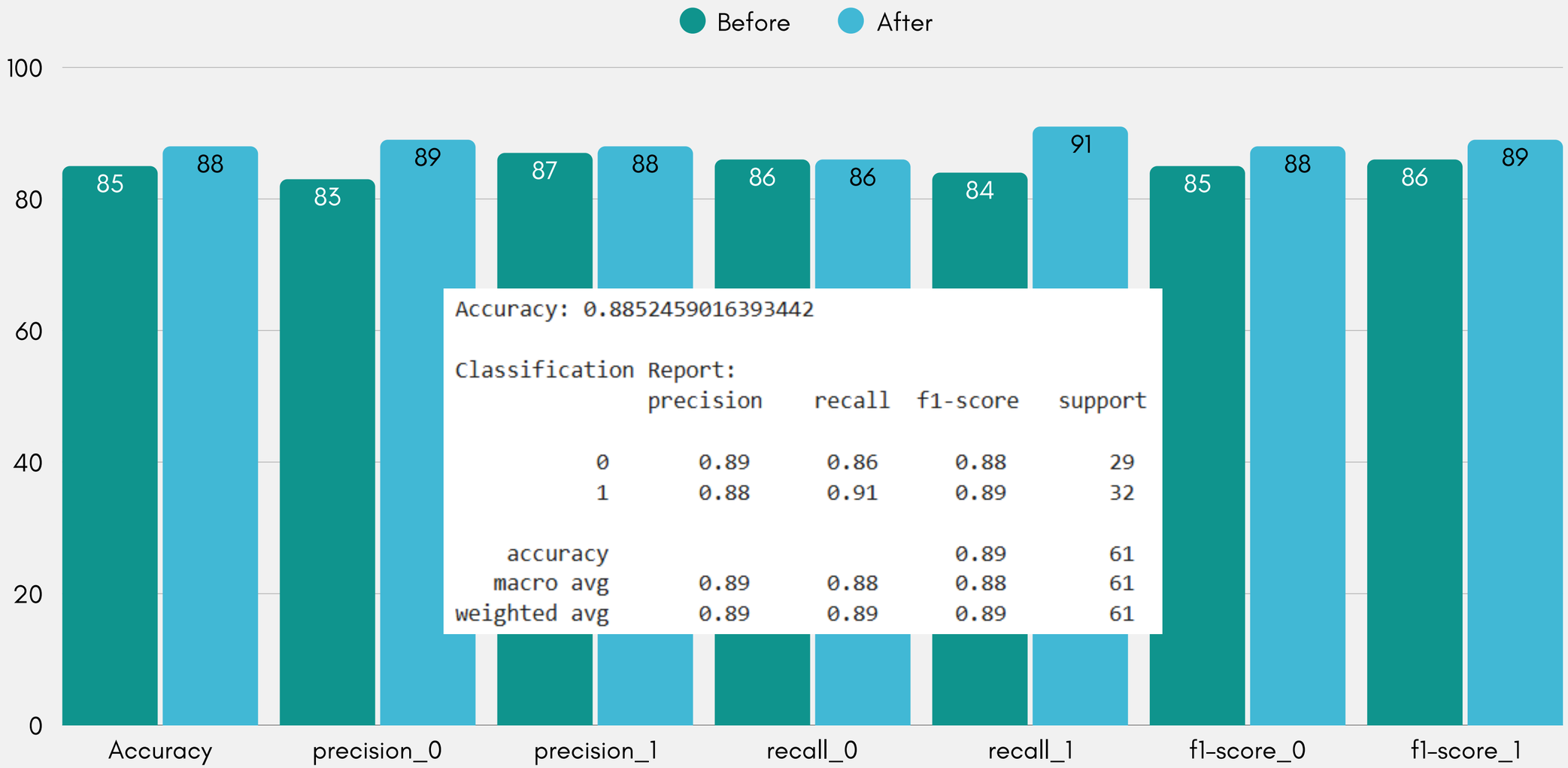
	feature	importance
9	oldpeak	0.128485
7	thalach	0.119725
11	ca	0.115533
2	cp	0.103792
12	thal	0.093300
0	age	0.092811
3	trestbps	0.077537
8	exang	0.075809
4	chol	0.074812
10	slope	0.051058
1	sex	0.035658
6	restecg	0.019782
5	fbs	0.011698

Feature Importance 값의 누적 합을 계산

=> 80% 이상의 중요도를 설명하는 변수까지 선택

독립 변수 13개 -> 8개

05. Feature Selection 로지스틱 회귀 분석 II



Accuracy: 0.8852459016393442

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.86	0.88	29
1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

06. 기대효과 및 활용방안 로지스틱 회귀 분석을 통한 심장 질환 발병 주요 원인 분석 결과

oldpeak

[운동으로 유발된 ST 분절 변화]
운동과 휴식 간 ST 분절의 차이 (mm)

cp

[흉통 유형]
환자가 경험한 흉통 유형을 네 가지로 구분
0 : 전형적인 협심증
1 : 비전형적인 협심증
2 : 비협심증성 통증
3 : 무증상

trestbps

[휴식 시 혈압]
입원 시 측정된 환자의 혈압 (mmHg)

thalach

[최대 심박수]
극한 운동 시 도달할 수 있는 최대 심박수 (bpm)

thal

[핵의학 스캔 결과]
심장 상태를 나타냄
1: 질환있음
2: 가역적 결함

exang

[운동 유발 협심증]
운동 중 협심증 발생 여부
1: 있음
0: 없음

ca

[주요 혈관 수]
혈관 조영술로 관찰된 주요 혈관의 수 (0~4)

age

[나이]
환자의 나이를 나타내는 연속형 데이터

06. 기대효과 및 활용방안

심장 질환 주요 변수의 정량적 분석

- 심장 질환의 위험 요인을 객관적으로 이해하는데 기여
- 특성 선택 기법으로 모델의 **단순화와 해석력 강화**

조기 진단과 예방을 위한 실질적 활용 가능성

- 도출된 주요 변수는 **조기 진단, 고위험군 선별**에 유용
- 생활습관 개선 및 건강 관리 방안 제공으로 **발병 예방**

의료 자원 활용의 효율성 증대

- 고위험군의 **효율적** 선별, 불필요한 검사 및 치료 최소화
- 조기 발견을 통한 치료 성공률 향상 -> **의료비 절감**

공중보건 정책 및 캠페인에의 기여

- 예방 캠페인 및 건강 교육 프로그램 등 **공중보건 정책 설계**에 활용
- 캠페인의 핵심 메시지, 고위험군 대상 **맞춤형 관리** 프로그램 개발

감사합니다