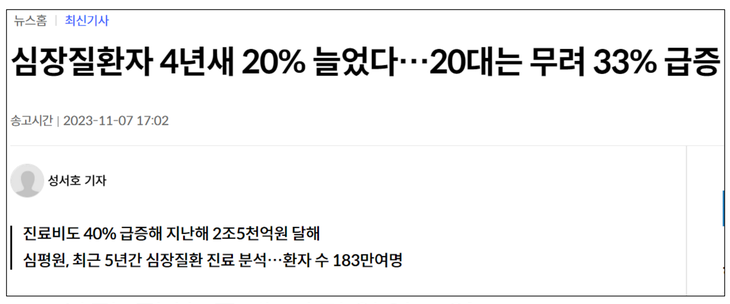
제2회 광운대학교 산업맞춤 단기직무능력 인증과정 매치업 심화과정 경진대회

로지스틱 회귀 분석을 활용한 심장 질환 주요 변수 예측

교과목명 / 의료빅데이터 분석 및 활용 실무(이석준)

김서경

INTRODUCTION


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

(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)

❖ 심장 질환 국내 사망원인 **2위**

❖ **젊은 층** 심장 실환 환자 가파른 **증가세**

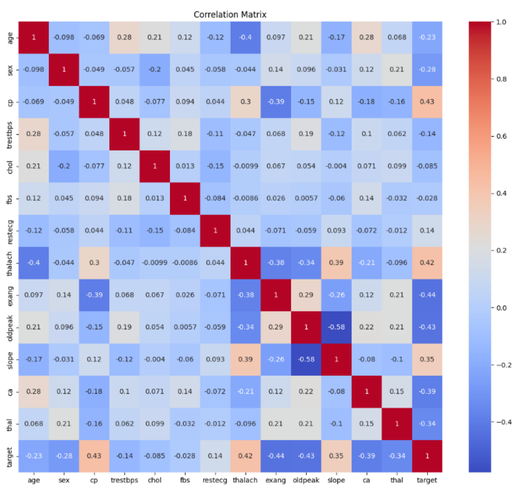
"로지스틱 회귀 분석을 통해 **심장 질환 발병 주요 원인을 분석**하자!"

DATA

❖ **14**개의 컬럼**, 304** 행, 종속변수 **target**

변수	설명
age	환자의 나이
sex	환자의 성별 (1: 남성, 2: 여성)
cp	흉통 유형(0: 전형적인 협심증, 1: 비전형적인 협심증, 2: 비협심증성 통증, 3: 무증상)
trestbps	병원 입원 시 혈압 (mmHg)
chol	혈중 총 콜레스테롤 수치 (mg/dL)
fbs	공복시 혈당이 120mg/dL을 초과하는 지 여부 (0: 정상, 1: 초과)
restecg	휴식 시 심전도 결과 (0: 정상, 1: ST-T파 이상, 2: 좌심실 비대 가능성)
thalach	최대 심박수 (bpm)
exang	운동 유발 협심증 여부 (1: 있음, 0: 없음)
oldpeak	운동 전후 ST 분절 차이 (mm)
slope	ST 분절 기울기 (0: 상승형, 1: 평형형, 2: 하강형)
ca	혈관 조영술로 관찰된 주요 혈관의 수 (0~4)
thal	핵의학 스캔 결과 (1: 질환 있음, 2: 가역적 결함)
target	**심장 질환 여부 (0: 질환 없음, 1: 질환 있음)**

EDA



❖ **상관 관계**

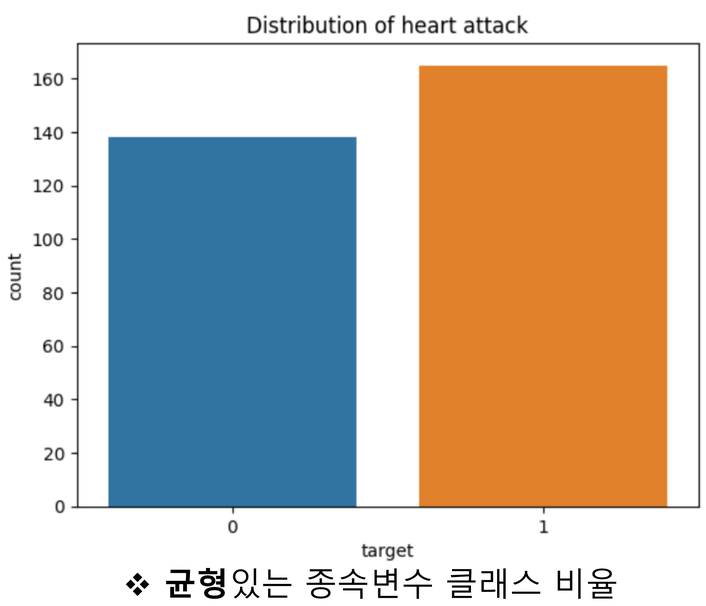
'target'과의 주요 관계		
양의 상관관계	cp(가슴통증)	0.43
	thalach(최대 심박수)	0.42
	slope(ST segment 기울기)	0.35
음의 상관관계	exang(운동 유발 협심증)	-0.44
	oldpeak(ST 분절기울기)	-0.43
	ca(주요 혈관 수)	-0.39

```
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
```

❖ **결측치 없음**



❖ **균형**있는 종속변수 클래스 비율

ANOVA

❖ 'target'과 **상관성이 높았던 세 범주형 변수**의 ANOVA , Post Hoc Test

One-Way ANOVA (Welch's)

	F	df1	df2	p
target	36.7	3	82.1	< .001

One-Way ANOVA (Welch's)

	F	df1	df2	p
target	28.0	2	54.9	< .001

One-Way ANOVA (Welch's)

	F	df1	df2	p
target	24.7	4	25.2	< .001

❖ ANOVA결과 모두 **높은 F값** -> 각 요인이 심장 질환에 미치는 영향 큼

❖ Post Hoc Test 결과 -> **범주 간 유의한 차이** 발견

cp = 0: 전형적인 협심증을 호소

slope = 2: ST 분절 기울기가 하강형

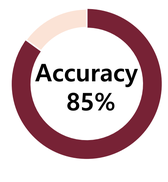
ca = 0: 주요 혈관의 수가 0

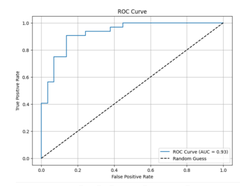
일 때, 다른 환자에 비해 심장 질환 발병률이 높을 수 있음

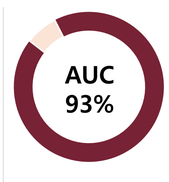
LOGISTIC REGRESSION

❖ CLASSIFICATION **REPORT**

CLASS	Precision	Recall	F1-Score	Support
Class 0	0.89	0.86	0.88	29
Class 1	0.88	0.91	0.89	32







❖ Cross-Validation (5-fold)시 **77% ~ 90%** 편차

교차 검증 정확도	평균 정확도
0.90163934	**83.83%**
0.85245902	
0.7704918	
0.81666667	
0.85	

❖ **유의하지 않은 coef 다수**

	coef	std err	z	P>|z|
const	-0.0330	0.206	-0.160	0.873
age	-0.0795	0.228	-0.349	0.727
sex	-0.8286	0.243	-3.411	0.001
cp	0.8862	0.219	4.048	0.000
trestbps	-0.3091	0.205	-1.509	0.131
chol	-0.1951	0.211	-0.924	0.356
fbs	0.1047	0.227	0.461	0.645
restecg	0.3106	0.208	1.493	0.135
thalach	0.4277	0.277	1.542	0.123
exang	-0.5434	0.216	-2.513	0.012
oldpeak	-0.7639	0.297	-2.576	0.010
slope	0.4700	0.243	1.931	0.053
ca	-0.8753	0.232	-3.773	0.000
thal	-0.6156	0.210	-2.928	0.003

FEATURE SELECTION

Variable	VIF
const	212.998773
age	1.443937
sex	1.231356
cp	1.397152
trestbps	1.180747
chol	1.152971
fbs	1.087698
restecg	1.066721
thalach	1.653567
exang	1.440147
oldpeak	1.744666
slope	1.662325
ca	1.290729
thal	1.191528
target	2.072754

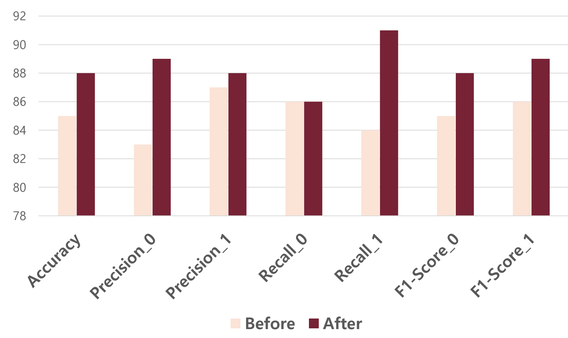
❖ **VIF ≒ 1**

Feature	Importance
oldpeak	0.128485
thalach	0.119725
ca	0.115533
cp	0.103792
thal	0.093300
age	0.092811
trestbps	0.077537
exang	0.075809
chol	0.074812
slope	0.051058
sex	0.035658
restecg	0.019782
fbs	0.011698

❖ **RandomForest**

❖ $N = \left\lfloor \frac{80}{\text{중요도}} \right\rfloor + 1$

❖ X변수 13 -> **8**



❖ 변수 선택 전 후 **성능 향상** 확인

OUTCOME

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

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

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