# 로지스틱 회귀 분석을 활용한 심장 질환 발병 여부에 따른 주요 요인 분석

정보융합학부 2020204051 김서경


1. 주제 선정
2. 데이터 개요
3. 데이터 EDA
   a. df.head()
   b. df.isnull().sum()
   c. plot
   d. Correlation Matrix
4. ANOVA 분석
   a. 연속형 변수 범주화 기준
   b. cp
   c. restecg
   d. slope
   e. ca
   f. thal
   g. age_group
   h. trestbps_group
   i. chol_group
5. 로지스틱 회귀 분석 I
   a. Accuracy
   b. Precision
   c. Recall
   d. F1-Score
   e. Support
   f. ROC-AUC
   g. Cross-Validation(5-Fold)
   h. 종합 해석
6. Feature Selection
   a. Logistic Regression Coefficients
   b. sm.Logit(y_train, X_train_const).fit()
   c. Variance Inflation Factor(VIF)
   d. RandomForestClassifier
7. 로지스틱 회귀 분석 II
   a. 정확도(Accuracy) 개선
   b. 정밀도(Precision)의 향상
   c. 재현율(Recall)의 개선
   d. F1-Score 상승
8. 결론
9. 기여점
   a. 심장 질환 주요 변수의 정량적 분석
   b. 조기 진단과 예방을 위한 실질적 활용 가능성
   c. 의료 자원 활용의 효율성 증대
   d. 공중보건 정책 및 캠페인에의 기여


## 1. 주제 선정

심장병은 국내에서 두 번째로 많은 사망 원인으로, 통계청(2024)에 따르면 2023년 한 해 동안 약 3만 3천 명이 심장질환으로 사망했다. 이는 국내 전체 사망자의 9.4%를 차지하며, 심장질환은 주요 공중보건 문제로 대두되고 있다. *(통계청, 2024 - '2023년 사망원인 통계')*

연합뉴스 보도에 따르면, 국내 심장질환 환자 수가 지난 4년간 약 20% 증가하여 2022년 기준 183만 명을 넘어섰다. 특히 20대 이하 젊은 층에서의 증가율이 두드러져, 같은 기간 동안 33% 급증한 것으로 나타났다. 이러한 통계는 심장질환이 더 이상 중·장년층에만 국한되지 않고, 젊은 층에서도 발병률이 높아지고 있음을 시사한다. (*연합뉴스, 2023 – '심장질환자 4년새 20% 늘었다... 20대는 무려 33% 급증'(성서호 기자)*)

뉴스홈 | 최신기사

**심장질환자 4년새 20% 늘었다…20대는 무려 33% 급증**

송고시간 | 2023-11-07 17:02

요약

성서호 기자

진료비도 40% 급증해 지난해 2조5천억원 달해
심평원, 최근 5년간 심장질환 진료 분석…환자 수 183만여명

(서울=연합뉴스) 성서호 기자 = 우리나라 사망 원인 2위인 심장질환 환자 수가 지난해 총 183만명을 넘어 진료비 2조5천억여원을 기록했다.

이는 4년 전보다 각각 20%, 40%가량 늘어난 것이다. 특히 20대 이하 젊은층에서 심장질환자의 비율이 크게 늘어난 것으로 분석됐다.

1147개

핫뉴스 +

中 후난성서 매장량 300t 거대 금광 발견…"총평가액 115조원"

비행기 같은 테이블석에 '통창' 개방감…미리 살펴본 판광버스

서천여고 운동장에서 조선시대 소

AI와 머신러닝 기술은 심장병 조기 진단과 예측에서 점차 중요한 도구로 자리 잡고 있다. 최근 분당서울대병원의 연구에 따르면, AI 기반 심혈관질환 예측 모델은 22만 명의 환자 데이터를 분석하여 75.1% 예측 정확도를 기록했다. 이 모델은 당뇨병력, 혈압, 콜레스테롤 수치 등 다양한 환자 데이터를 활용하여 개인별 심장병 발병 위험을 정확히 평가할 수 있었다. *(AITimes, 2021 – "국내 연구진, AI 머신러닝 기반 '한국형 심혈관질환 예측 모델' 개발"(최광민 기자))*

따라서, 데이터를 로지스틱 회귀분석을 통해 심장병을 조기에 예측하고 치료하는데 기여할 수 있을 것이다.

## 2. 데이터 개요

데이터는 kaggle의 'Health care: Heart attack possibility' data를 활용했다. 이 데이터는 심장병 연구에 초점을 맞춘 데이터로서 14개의 주요 속성을 추린 데이터이다. 데이터의 출처는 Cleveland 데이터베이스로, 이는 머신러닝(Machine Learning) 연구자들이 사용해 온 유일한 데이터셋이며, 심장병 예측 모델 연구에서 주로 활용되었다. 종속변수가 될 target 컬럼은 심장병 발생 유무를 나타낸다. 값은 정수형이며 '0' 일 때는 심장질환 가능성이 낮거나 없음을 나타내고, '1' 일땐 심장질환 가능성이 높음을 나타낸다. 데이터셋은 심장 질환의 조기 진단 및 예방을 위한 의학적 연구와 머신러닝 모델 개발에 중요한 역할을 한다. 특히, 환자의 상태를 빠르게 분류하고 적절한 치료를 지원하는

데 유용할 수 있다.

### 3. 데이터 EDA

#### a. df.head

| | 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 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

데이터 모양을 보면, 모든 데이터가 수치형 데이터로 이루어져있음을 볼 수 있다. 종속변수를 포함해서 총 14개의 컬럼으로 이루어져있고, 304개의 데이터 행이 존재한다. 주요 변수 설명은 아래와 같다.

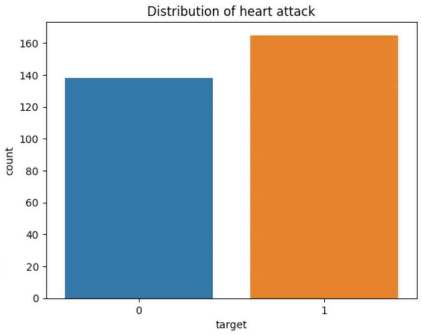
- age (나이) : 환자의 나이를 나타내는 연속형 데이터
- sex (성별) : 환자의 성별을 나타내며, 1은 남성, 2는 여성
- cp (흉통 유형) : 환자가 경험한 흉통 유형을 네 가지로 구분
  - 0 : 전형적인 협심증
  - 1 : 비전형적인 협심증
  - 2 : 비협심증성 통증
  - 3 : 무증상
- trestbps (휴식 시 혈압) : 병원 입원 시 측정된 환자의 혈압 (단위: mmHg)
- chol (콜레스테롤) : 혈액 내 총 콜레스테롤 수치 (단위: mg/dL)
- fbs (공복 혈당) : 공복 시 혈당이 120mg/dL를 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)

#### b. df.isnull().sum()

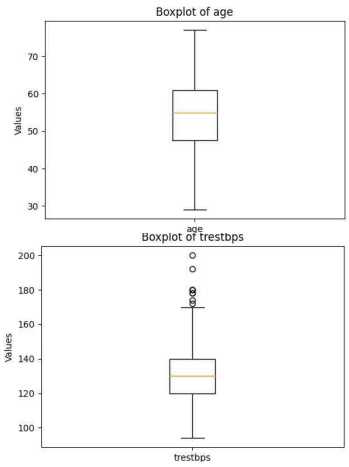
```
[2]:

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
```

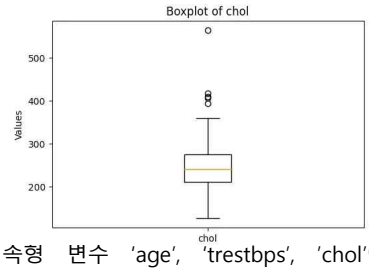
데이터 내의 결측을 확인해본 결과 결측값은 존재하지 않았다.

#### c. plot



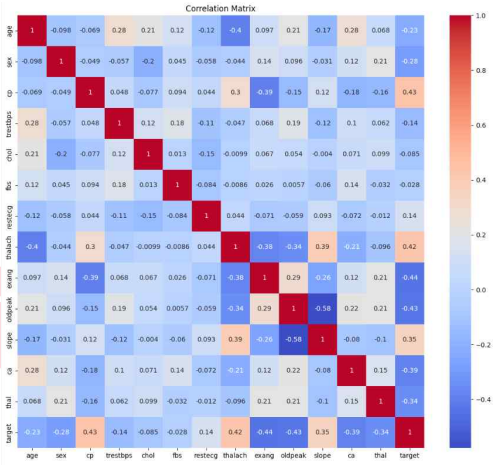

종속변수 'taget'에 대한 barplot을 도출해보니, 해당 데이터셋에는 비슷한 비율로 심장질환이 발생한 사람과 발생하지 않은 사람의 데이터가 있음을 알수 있다. 이는 오버샘플링이나 언더샘플링과 같은 데이터 불균형 문제를 띄지 않는 데이터임을 시사한다.






연속형 변수 'age', 'trestbps', 'chol'에 대해 boxplot을 그려봤을때, 0개에서 6개 정도의 데이터가 3QR 안에 위치하지 않았다. 하지만 데이터셋 자체의 데이터행 개수가 304개로 13개의 컬럼을 활용해 분석하기에 많은 양의 데이터가 아니기때문에 데이터를 삭제하지 않기로 결정했다.

**d. Correlation Matrix**




이 행렬은 각 변수 간의 상관계수를 시각적으로 나타낸 것이다. 상관계수의 값 범위는 -1에서 1까지이며 1에 가까운 값은 강한 양의 상관관계 (한 변수가 증가하면 다른 변수도 증가), -1에 가까운 값은 강한 음의 상관관계 (한 변수가 증가하면 다른 변수는 감소), 0에 가까운 값은 거의 상관관계가 없음을 의미한다. 빨간색은 높은 양의 상관관계를, 파란색은 높은 음의 상관관계를, 흰색이나 밝은 색은 상관관계가 거의 없음을 보여준다.

먼저, 'cp'(가슴 통증)과 'target'의 상관계수는 0.43으로, 양의 상관관계를 나타낸다. 이는 가슴 통증이 증가할수록 'target' 값도 증가하는 경향이 있음을 시사한다. 다음으로, 'thalach'(최대 심박수)와 'target'의 상관계수는 0.42로, 최대 심박수가 높을수록 'target' 값이 높아지는 경향을 보여준다.

반대로, 'exang'(운동으로 유발된 협심증)과 'target'은 -0.44로 음의 상관관계를 가지며, 이는 'exang' 값이 높아질수록 'target' 값은 낮아진다는 것을 의미한다.

또한, 'oldpeak'(ST depression)과 'target'의 상관계수는 -0.43으로 음의 상관관계를 나타낸다. 이는 ST depression 값이 증가할수록 'target' 값이 낮아지는 경향이 있음을 보여준다. 'slope'(ST segment의 기울기)는 0.35로 'target'과 양의 상관관계를 가지며, 이 값이 높아질수록 'target' 값이 증가하는 경향을 보인다. 반면, 'ca'(주요 혈관 수)는 -0.39로 'target'과 음의 상관관계를 가지며, 혈관 수가 많아질수록 'target' 값이 낮아지는 경향이 있다.

그 외에도, 변수 간 주목할 만한 상관관계로는 'thalach'와 'exang'의 상관계수가 -0.38로 나타나며, 이는 최대 심박수와 운동 유발 협심증이 음의 상관관계를 가짐을 보여준다. 또한, 'oldpeak'와 'slope'는 -0.58로 강한 음의 상관관계를 가지며, ST depression 값이 증가할수록 ST segment 기울기가 감소하는 경향을 나타낸다.

종합적으로, 'cp', 'thalach', 'exang', 'oldpeak', 'slope', 'ca'가 'target'과 비교적 높은 상관관계를 보이는 변수들로 확인되며, 데이터 분석 및 예측 모델 개발 시 이 변수들을 중심으로 고려하는 것이 효과적일 것으로 보인다.

## 4. ANOVA 분석

연속형 변수에 대해 종속 변수(이진 클래스) 와의 분산 차이를 통계적으로 검정하여 변수 중요도를 판단하기 위해 ANOVA 분석을 시행했다.

**a. 연속형 변수 범주화 기준**

| 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 | |

범주형 변수 cp, restecg, slope, ca, thal 와 구간별로 범주화한 연속형 변수인 age, trestbps, chol에 대해서 ANOVA를 진행했다.

(단위: 명, %)

| 구분 | 전체 | 9세 이하 | 10대 | 20대 | 30대 | 40대 | 50대 | 60대 | 70대 | 80대 이상 |
|---|---|---|---|---|---|---|---|---|---|---|
| 계 | 942,017 | 212 | 837 | 4,341 | 12,364 | 50,985 | 168,129 | **289,873** | 274,739 | 140,537 |
| | (100) | (0.0) | (0.1) | (0.5) | (1.3) | (5.4) | (17.8) | **(30.8)** | (29.2) | (14.9) |
| 남성 | 584,953 | 151 | 536 | 2,799 | 9,280 | 39,987 | 122,737 | **191,802** | 155,603 | 62,058 |
| | (100) | (0.0) | (0.1) | (0.5) | (1.6) | (6.8) | (21.0) | **(32.8)** | (26.6) | (10.6) |
| 여성 | 357,064 | 61 | 301 | 1,542 | 3,084 | 10,998 | 45,392 | 98,071 | **119,136** | 78,479 |
| | (100) | (0.0) | (0.1) | (0.4) | (0.9) | (3.1) | (12.7) | (27.5) | **(33.4)** | (22.0) |

2019년 '허혈성 심장질환' 연령대별 / 성별 진료인원 (건보공단 제공)

나이 변수 'age'에 대해서는 건보공단에서 공개한 허혈성 심장질환 연령대별 소계와 데이터의 분포

를 기준으로, '60대 미만', '60-69세', '70세 이상'으로 범주화 하였다.

〈표 2018 고혈압 진료지침에 따른 혈압의 분류〉

| 혈압 분류 | | 수축기 혈압(mmHg) | | 이완기 혈압(mmHg) |
|---|---|---|---|---|
| 정상 혈압* | | <120 | 그리고 | <80 |
| 주의 혈압 | | 120~129 | 그리고 | <80 |
| 고혈압 전단계 | | 130~139 | 또는 | 80~89 |
| 고혈압 | 1기 | 140~159 | 또는 | 90~99 |
| | 2기 | ≥160 | 또는 | ≥100 |
| 수축기 단독 고혈압 | | ≥140 | 그리고 | <90 |

*심뇌혈관 질환의 발생 위험이 가장 낮은 최적혈압.

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

〈그림 이상지질혈증 분류 기준〉


*출처 : 한국지질·동맥경화학회 진료지침위원회 (2022). 이상지질혈증 진료지침 5판. 한국지질·동맥경화학회.

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

혈압 변수 trestbps는 고혈압 진료 지침에 따라 4단계로 범주화했고, 콜레스테롤 수치를 나타내는 변수 chol은 한국지질동맥강화학회 진료지침위원회에서 발표한 총콜레스테롤의 분류 기준을 차용해 범주화했다.

**b. cp**

**One-Way ANOVA**

cp

One-Way ANOVA (Welch's)

| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 36.7 | 3 | 82.1 | < .001 |

**Post Hoc Tests**

Games-Howell Post-Hoc Test – target

| | | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| 0 | Mean difference | — | -0.547 | -0.5204 | -0.4229 |
| | p-value | — | < .001 | < .001 | 0.002 |
| 1 | Mean difference | | — | 0.0269 | 0.1243 |
| | p-value | | — | 0.981 | 0.688 |
| 2 | Mean difference | | | — | 0.0975 |
| | p-value | | | — | 0.801 |
| 3 | Mean difference | | | | — |
| | p-value | | | | — |

F 값이 높다는 것은 cp(흉통 유형)이 target(심장 질환 여부)에 상당한 영향을 미친다는 것을 나타낸다. p-값이 0.001보다 작아 통계적으로 유의미한 결과를 보여주며 이는 흉통 유형에 따라 심장 질환 여부에 차이가 있음을 의미한다. 사후 분석을 통해 그룹 간의 평균 차이를 구체적으로 비교하였다.

- 집단 0 (전형적인 협심증)
집단 1, 2, 3과 비교 시, 평균 차이가 모두 유의미한 것으로 나타난다.
이는 전형적인 협심증을 경험한 환자가 다른 유형의 흉통을 경험한 환자들보다 심장 질환 발병 가능성이 더 크거나 낮다는 것을 의미한다.

- 집단 1 (비전형적인 협심증)
집단 2와의 평균 차이는 유의하지 않다.
집단 3과의 평균 차이도 유의하지 않다.
이는 비전형적인 협심증 환자들이 다른 흉통 유형과 비교했을 때 심장 질환 발병률에 큰 차이가 없음을 나타낸다.

- 집단 2 (비협심증성 통증)
집단 3과 비교 시 평균 차이가 유의하지 않다.
이는 비협심증성 통증과 무증상 환자들이 심장 질환 발병 가능성이 비슷함을 의미한다.

흉통 유형에 따른 심장 질환 발병률 차이는 주로 전형적인 협심증 환자와 다른 집단(비전형적, 비협심증성 통증, 무증상) 간에서 유의미하게 나타난다. 전형적인 협심증은 심장 질환과 가장 강한 연관성을 가진 흉통 유형으로 보이며, 나머지 유형들(특히 비전형적 협심증, 비협심증성 통증, 무증상) 간에는 뚜렷한 차이가 관찰되지 않는다. 따라서 전형적인 협심증을 호소하는 환자는 심장 질환 발병 가능성이 높기 때문에 조기 진단과 적극적인 치료가 필요하다.

**c. restecg**

restecg

One-Way ANOVA (Welch's)

| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 4.73 | 2 | 8.12 | 0.043 |

**Post Hoc Tests**

Games-Howell Post-Hoc Test – target

| | | 0 | 1 | 2 |
|---|---|---|---|---|
| 0 | Mean difference | — | -0.169 | 0.213 |
| | p-value | — | 0.009 | 0.707 |
| 1 | Mean difference | | — | 0.382 |
| | p-value | | — | 0.399 |
| 2 | Mean difference | | | — |
| | p-value | | | — |

p-값이 0.05보다 작으므로, restecg(휴식 시 심전도 상태) 그룹 간에 심장 질환 여부의 평균 차이가 통계적으로 유의미하다고 해석되며 이는 심전도 상태에 따라 심장 질환 발병률이 다를 가능성을 시사한다.

- 집단 0 (정상)
집단 1(ST-T파 이상)과 비교 시, 평균 차이가 유의미하다.
집단 2(좌심실 비대 가능성)과 비교 시, 평균 차이가 유의하지 않다.
이는 정상 심전도를 가진 환자와 ST-T파 이상을 보이는 환자 간 심장 질환 발병률에 유의미한 차이가 있음을 나타낸다.

- 집단 1 (ST-T파 이상)
집단 2(좌심실 비대 가능성)과 비교 시, 평균 차이가 유의하지 않다.
이는 ST-T파 이상과 좌심실 비대 가능성을 보이는 환자 간 심장 질환 발병률에 큰 차이가 없음을 나타낸다.

- 집단 2 (좌심실 비대 가능성)
집단 0(정상) 또는 집단 1(ST-T파 이상)과 비교 시, 평균 차이가 유의하지 않다.

restecg 변수의 세 그룹 중 정상 심전도(집단 0)와 ST-T파 이상(집단 1) 간에만 유의미한 차이가 존재한다. 이는 ST-T파 이상이 심장 질환 발병률과 더 높은 연관성을 가질 가능성을 시사한다. ST-T파 이상(집단 1)을 보이는 환자는 심장 질환 발병 가능성이 높으므로 추가적인 검진 및 치료가 필요하다. 정상 심전도(집단 0)를 보이는 환자는 상대적으로 심장 질환 발병 가능성이 낮을 가능성이 크다.

**d. slope**

**One-Way ANOVA**

slope

One-Way ANOVA (Welch's)

| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 28.0 | 2 | 54.9 | < .001 |

**Post Hoc Tests**

Games-Howell Post-Hoc Test – target

| | | 0 | 1 | 2 |
|---|---|---|---|---|
| 0 | Mean difference | — | 0.0786 | -0.325 |
| | p-value | — | 0.785 | 0.026 |
| 1 | Mean difference | | — | -0.404 |
| | p-value | | — | < .001 |
| 2 | Mean difference | | | — |
| | p-value | | | — |

F 값이 높게 나타나 slope(ST 분절 기울기)가 target(심장 질환 여부)에 미치는 영향이 상당히 크다는 것을 나타낸다. p-값이 0.001보다 작으므로, Slope 그룹 간에 심장 질환 여부의 평균 차이가 통계적으로 유의미하다고 해석되며 이는 ST 분절 기울기 형태에 따라 심장 질환 발병률이 달라질 가능성을 시사한다.

- 집단 0 (상승형)
집단 1(평형형)과 비교 시 평균 차이가 유의하지 않다.
집단 2(하강형)과 비교 시 평균 차이가 유의미하다.
이는 상승형 기울기를 가진 환자와 하강형 기울기를 가진 환자 간 심장 질환 발병률에 유의미한 차이가 있음을 나타낸다.

- 집단 1 (평형형)
집단 2(하강형)과 비교 시 평균 차이가 유의미하다.
이는 평형형 기울기를 가진 환자와 하강형 기울기를 가진 환자 간 심장 질환 발병률에 유의미한 차이가 있음을 나타낸다.

- 집단 2 (하강형)
집단 0(상승형) 및 집단 1(평형형)과 비교하여 심

장 질환 발병률에 유의미한 차이를 보인다.
하강형 기울기를 가진 환자가 다른 기울기 유형을 가진 환자들보다 심장 질환 발병률이 더 높음을 시사한다.

ST 분절 기울기 형태는 심장 질환 발병률과 밀접한 연관이 있다. 특히, 하강형 기울기(집단 2)는 다른 기울기 유형(상승형, 평형형)과 비교하여 심장 질환 발병률이 유의미하게 높다. 따라서 하강형 기울기(집단 2)를 보이는 환자는 심장 질환 발병 가능성이 가장 높으므로, 조기 진단과 적극적인 치료가 필요하다.

**e. ca**

**One-Way ANOVA**

ca

One-Way ANOVA (Welch's)

| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 24.7 | 4 | 25.2 | < .001 |

**Post Hoc Tests**

Games-Howell Post-Hoc Test – target

| | | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| 0 | Mean difference | — | 0.420 | 0.559 | 0.5929 | -0.0571 |
| | p-value | — | < .001 | < .001 | < .001 | 0.998 |
| 1 | Mean difference | | — | 0.139 | 0.1731 | -0.4769 |
| | p-value | | — | 0.498 | 0.434 | 0.290 |
| 2 | Mean difference | | | — | 0.0342 | -0.6158 |
| | p-value | | | — | 0.997 | 0.148 |
| 3 | Mean difference | | | | — | -0.6500 |
| | p-value | | | | — | 0.126 |
| 4 | Mean difference | | | | | — |
| | p-value | | | | | — |

F 값이 높게 나타나 ca(주요 혈관 수)가 target(심장 질환 여부)에 미치는 영향이 크다는 것을 나타낸다. p-값이 0.001보다 작으므로, ca 그룹 간에 심장 질환 여부의 평균 차이가 통계적으로 유의미하다고 해석되며 이는 주요 혈관 수에 따라 심장 질환 발병률이 달라질 가능성을 시사한다.

- 집단 0과의 비교
집단 1, 2, 3과 비교 시 평균 차이가 모두 유의미하다.
집단 4와의 평균 차이는 유의하지 않다.
이는 혈관 수가 0인 환자가 혈관 수가 1, 2, 3인 환자와 심장 질환 발병률에서 유의미한 차이를 보이는 반면, 혈관 수가 4인 환자와는 차이가 없음을 나타낸다.

- 집단 1과의 비교
집단 2, 3, 4와의 평균 차이는 모두 유의하지 않다.
이는 혈관 수가 1인 환자가 다른 혈관 수 그룹(2, 3, 4)과 심장 질환 발병률에서 유의미한 차이를 보이지 않음을 나타낸다.

- 집단 2와의 비교
집단 3과의 평균 차이는 유의하지 않다.
집단 4와의 평균 차이는 유의하지 않다.
이는 혈관 수가 2인 환자가 혈관 수가 3 또는 4인 환자와 심장 질환 발병률에서 유의미한 차이를 보이지 않음을 나타낸다.

- 집단 3과 집단 4의 비교
평균 차이는 유의하지 않다.
이는 혈관 수가 3인 환자와 4인 환자가 심장 질환 발병률에서 유의미한 차이를 보이지 않음을 나타낸다.

집단 간 유의미한 차이는 주로 혈관 수가 0인 환자와 혈관 수가 1, 2, 3인 환자 간에서 나타난다. 혈관 수가 4인 환자는 다른 그룹(특히 혈관 수 0)과 심장 질환 발병률에서 유의미한 차이를 보이지 않는다. 따라서 혈관 수가 0인 환자는 다른 혈관 수 그룹(특히 1, 2, 3)에 비해 심장 질환 발병률이 유의미하게 높을 가능성이 있으므로, 추가적인 진단과 조기 치료가 필요하며 혈관 수가 4인 환자는 다른 그룹과 유사한 발병률을 보여, 추가적인 검토가 필요하지 않을 수 있다.

**f. thal**

**One-Way ANOVA**

thal

One-Way ANOVA (Welch's)

| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 30.7 | 3 | 4.75 | 0.002 |

**Post Hoc Tests**

Games-Howell Post-Hoc Test – target

| | | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| 0 | Mean difference | — | 0.167 | -0.283 | 0.2607 |
| | p-value | — | NaN | NaN | NaN |
| 1 | Mean difference | | — | -0.450 | 0.0940 |
| | p-value | | — | 0.006 | 0.864 |
| 2 | Mean difference | | | — | 0.5438 |
| | p-value | | | — | < .001 |
| 3 | Mean difference | | | | — |
| | p-value | | | | — |

F 값이 높게 나타나 thal(핵의학 스캔 결과, 심장 상태; 1: 질환 있음 , 2: 가역적 결함)가 target(심

장 질환 여부)에 미치는 영향이 크다는 것을 나타낸다. p-값이 0.05보다 작으므로, Thal 그룹 간에 심장 질환 여부의 평균 차이가 통계적으로 유의미하다고 해석되며 이는 핵의학 스캔 결과에 따라 심장 질환 발병률이 달라질 가능성을 시사한다.

- 집단 1과의 비교
집단 2와 비교 시, 평균 차이는 유의미하다.
집단 3과 비교 시, 평균 차이는 유의하지 않다.
이는 집단 1과 집단 2 간 심장 질환 발병률에 유의미한 차이가 있으며, 집단 1과 집단 3 간에는 유의미한 차이가 없음을 나타낸다.

- 집단 2와의 비교
집단 3과 비교 시, 평균 차이는 유의미하다.
이는 집단 2와 집단 3 간 심장 질환 발병률에 유의미한 차이가 있음을 나타낸다.

집단 간 유의미한 차이는 주로 집단 1과 집단 2, 그리고 집단 2와 집단 3 간에서 관찰된다. 집단 2와 집단 3 간에는 심장 질환 발병률에서 뚜렷한 차이가 나타나므로, 두 그룹의 차이를 중심으로 추가적인 분석이 필요하다. 집단 1과 집단 2 간에도 유의미한 차이가 있으므로, 집단 1 환자와 집단 2 환자의 특징을 비교하여 추가적인 분석이 필요하다.

**g. age_group**

**One-Way ANOVA**

age_group

One-Way ANOVA (Welch's)

| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 4.55 | 2 | 23.9 | 0.021 |

**Post Hoc Tests**

Games-Howell Post-Hoc Test – target

| | | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | Mean difference | — | 0.196 | -0.00376 |
| | p-value | — | 0.008 | 1.000 |
| 2 | Mean difference | | — | -0.20000 |
| | p-value | | — | 0.499 |
| 3 | Mean difference | | | — |
| | p-value | | | — |

p-값이 0.05보다 작으므로, 연령대 그룹 간에 심장 질환 여부의 평균 차이가 통계적으로 유의미하며 이는 연령대에 따라 심장 질환 발병률이 달라질 가능성을 시사한다.

- 60세 미만과의 비교
60–69세와의 평균 차이는 유의미하다. 이는 60세 미만과 60–69세 그룹 간 심장 질환 발병률에서 유의미한 차이가 있음을 의미한다.
70–79세와의 평균 차이는 유의미하지 않다. 이는 60세 미만과 70–79세 그룹 간 심장 질환 발병률에 차이가 없음을 의미한다.

- 60–69세와의 비교
70–79세와의 평균 차이는 유의미하지 않다. 이는 60–69세와 70–79세 그룹 간 심장 질환 발병률에 차이가 없음을 의미한다.

연령대 그룹 간 유의미한 차이는 60세 미만(그룹 1)과 60–69세(그룹 2) 간에서만 관찰되며 70–79세와 다른 연령대 그룹 간에는 심장 질환 발병률의 차이가 유의미하지 않다. 60세 미만(그룹 1)과 60–69세(그룹 2) 간 차이가 관찰되는 점을 통해, 60세를 경계로 심장 질환 발병률에 변화가 있을 가능성을 고려해야 한다.

**h. trestbps_group**

**One-Way ANOVA**

trestbps_group

One-Way ANOVA (Welch's)

| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 2.28 | 3 | 92.0 | 0.084 |

trestbps_group 컬럼의 경우, ANOVA 분석 결과가 0.084로 통계적으로 유의하지 않았다.

**i. chol_group**

**One-Way ANOVA**

chol_group

One-Way ANOVA (Welch's)

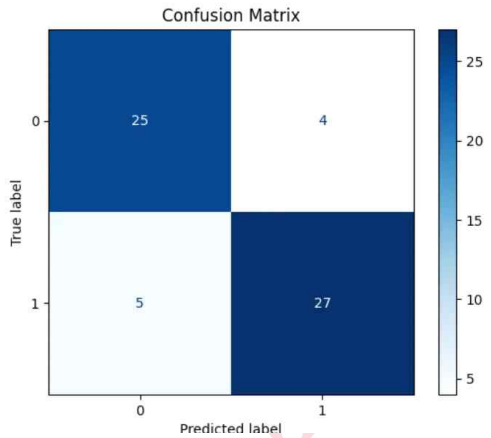
| | F | df1 | df2 | p |
|---|---|---|---|---|
| target | 1.88 | 2 | 131 | 0.157 |

chol_group 컬럼의 경우도, ANOVA 분석 결과가 0.157로 통계적으로 유의하지 않았다.

### 5. 로지스틱 회귀 분석 I

우선, 종속변수인 'target' 데이터를 데이터셋과 분리해 X와 y로 구분하였고, 8:2의 비율로 test set과 train set을 랜덤샘플링을 통해 가졌다. 로지스틱 회귀는 가중치(weight)에 민감하므로, StandardScaler를 사용하여 데이터를 표준화한 후, 로지스틱 회귀 분석 모델에 적용했다. 회귀 분석 결과 평가지표 값과, 혼동 행렬 값은 아래와 같았다.

```
Accuracy: 0.8524590163934426

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.86      0.85        29
           1       0.87      0.84      0.86        32

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61
```




#### a. Accuracy

Accuracy는 전체 샘플 중에서 모델이 정확히 예측한 샘플의 비율이다. 모델이 전체 샘플의 약 85.25%를 올바르게 분류했음을 알 수 있다.

#### b. Precision

Precision은 모델이 양성(1, positive)로 예측한 것 중에서 실제로 양성인 데이터의 비율이다. 위 평가값을 보면 로지스틱 회귀 분석 모델이 클래스 0으로 예측한 데이터 중 약 83%가 실제로 클래스 0이었고, 모델이 클래스 1로 예측한 데이터 중 약 87%가 실제로 클래스 1이었음을 알 수 있다. 높은 정밀도를 띄고 있기 때문에 잘못된 양성 예측이 적음을 알 수 있다.

#### c. Recall

Recall은 모델이 실제 양성 데이터를 얼마나 잘 예측했는지를 나타낸다. 실제 클래스 0인 데이터 중 약 86%를 모델이 정확히 예측했고, 실제 클래스 1인 데이터 중 약 84%를 모델이 정확히 예측했다. 높은 재현율을 통해 잘못된 음성 예측이 적었음을 알수 있다.

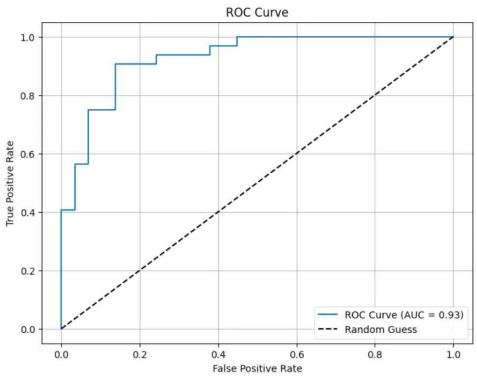
#### d. F1-Score

F1-Score는 Precision과 Recall의 조화 평균으로, 두 지표 간의 균형을 평가한다. 평가 결과, 클래스 0에 대해서는 0.85, 클래스 1에 대해서는 0.86으로 두 지표간 균형을 잘 이룸을 알 수있다.

#### e. Support

Support는 각 클래스에 속한 샘플의 개수를 나타내는데, 클래스 0과 1에 대해서 각각 29개 32개로 클래스별 샘플 수의 큰 차이가 없어 모델 평가가 신뢰할 만 하다고 해석할 수 있다.

#### f. ROC-AUC




이진 분류 모델의 전반적인 성능을 평가하기 위해 ROC-AUC 점수도 확인해보았다. X축은 (False Positive Rate, FPR) 잘못된 양성 예측 비율을 나타내고, Y축은 (True Positive Rate, TPR) 올바른 양성 예측 비율, 즉 재현율(Recall)을 나타낸다. ROC Curve는 분류 임계값(threshold)을 조정하면서 FPR과 TPR의 관계를 나타낸 곡선이다. ROC Curve는 왼쪽 위로 치우쳐져 있어, FPR이 낮은 상태에서도 높은 TPR을 유지하고 있음을 보여준다. 이는 모델이 양성과 음성을 잘 구분하고 있음을 의미한다. AUC는 ROC Curve 아래 면적을 의미하며, 모델의 분류 성능을 하나의 값으로 나타낸다.

AUC의 값이 1.0이면 완벽한 분류 (모델 성능이 매우 좋음), 0.5면 무작위 추측 (랜덤 예측 수준), < 0.5면 분류 성능이 나쁜 경우 (잘못된 예측)로 해석한다. 이 모델은 AUC 값이 0.93으로 매우 높은 수준이고 이 값은 모델이 무작위 추측보다 훨씬 뛰어난 성능을 보이며, 93%의 확률로 양성과 음성을 올바르게 구분할 수 있음을 의미한다.

### g. Cross-Validation (5-Fold)

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

교차 검증(Cross-Validation)은 데이터를 여러 번 나누어 모델을 평가하는 방법으로, 모델의 일반화 성능을 더 신뢰성 있게 평가하는 데 사용된다. 데이터를 훈련 세트와 테스트 세트로 나누는 과정에서 발생할 수 있는 편향을 줄이고, 모델이 새로운 데이터에서도 잘 작동할 가능성을 측정한다. StratifiedKFold를 사용해 클래스 비율을 유지한 상태로 데이터를 나누고 데이터를 5개의 Fold로 나누어 5번 학습 및 검증을 시행했다. 전반적으로 모델의 성능은 평균적으로 83.83%을 띄어 안정적이며, 일반화 가능성이 높은 것으로 평가된다.

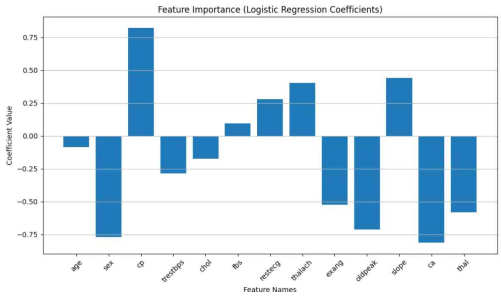
### h. 종합 해석

클래스 0과 클래스 1 모두에서 Precision, Recall, F1-Score가 고르게 높게 나타나, 모델이 두 클래스에 대해 균형 잡힌 성능을 보이고 있다. 클래스 간 샘플 수가 큰 차이가 없으므로, Macro Avg와 Weighted Avg의 값이 비슷하게 나타났다. 이 모델은 클래스 0과 클래스 1 모두에서 잘 작동하며, 오분류의 비율이 낮다. AUC 0.93을 달성하며, 높은 분류 성능을 보이는 우수한 모델임과 동시에, ROC Curve의 모양은 낮은 False Positive Rate에서도 높은 True Positive Rate을 유지하고 있어, 양성과 음성을 효과적으로 구분하고 있다고 할 수 있다. 모델의 성능이 전반적으로 뛰어나며, 만약 특정 상황에서 False Positive 또는 False Negative를 더 중요하게 다뤄야 한다면, ROC Curve를 참고하여 적절한 임계값을 설정할 수 있다. 예를 들어, 의료 데이터에서는 False Negative (질환을 놓침)를 줄이는 것이 중요하므로 TPR(Recall)을 높이는 방향으로 임계값을 조정할 수 있다.

다만, 교차 검증 과정에서 77%와 90%로 예측 간 편차가 큼을 알수 있었기 때문에 Logistic Regression Coefficients 를 이용해 특성 중요도를 파악하고, VIF 계수를 통한 다중공선성 확인, RandomForest를 활용한 Feature Selection을 활용해 각 독립 변수가 종속 변수에 미치는 영향을 확인하고 로지스틱 회귀 분석의 성능을 향상시키고자 한다.

## 6. Feature Selection

### a. Logistic Regression Coefficients




모델이 사용하는 주요 변수를 식별하기 위해서 Logistic Regression Coefficients를 시각화하였다. 각 변수의 계수(Coefficient) 값은 해당 변수가 종속 변수'target'에 미치는 영향을 나타낸다. x축은 독립 변수의 이름을 나타내고, y축은 각 변수의 회귀 계수 값이다. 이 값은 종속 변수에 대한 해당 변수의 영향을 나타낸다.

계수가 양수면, 해당 변수가 증가하면 종속 변수의 예측 확률이 증가하고, 음수 계수면, 해당 변수가 증가하면 종속 변수의 예측 확률이 감소한다.

계수의 절대값이 클수록 변수의 중요도가 높음을 알수있다. 반대로, 계수의 절대값이 작다면 해당 변수는 종속 변수에 상대적으로 적은 영향을 미친다고 볼 수 있다. 예를 들어, 가슴 통증을 나타내는 변수인 'cp'는 가장 높은 양의 계수를 가진다. 이는 cp가 증가할수록(가슴 통증 강도가 높아질수록) , 종속 변수(심혈관 질환 발생 확률 등)가 증가함을 의미한다.

### b. sm.Logit(y_train, X_train_const).fit()

p-value로 Logistic Regression Coefficients가 유의미한지 확인하기 위해 statsmodels를 사용한 로지스틱 회귀 모델에 적합시켜 보았다. 그 결과 위 그림과 같은 해석이 나왔다. P > | z | 는 Logistic Regression Coefficients가 통계적으로 유의한 해석이 나왔는 지를 확인하는 값이다. 0.05 미만의 p-value를 갖는 독립 변수가 총 13개 중에서 6개로 'sex', 'cp', 'exang', 'oldpeak', 'ca', 'thal' 이었다. 절반 정도의 변수가 낮은 통계적 유의성을 가지고 있었기 때문에, VIF 를 통해 다중공선성의 문제는 없는지 확인하고자 하였다.

```
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
==============================================================================
```

**c. 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 있다. 계산한 Logistic Regression Coefficients가 절반 이상의 독립변수에서 유의하지 않은 결과값이 나왔기 때문에 변수간 분산팽창지수(Variance Inflation Factor,VIF)를 통해 다중공선성을 확인해보았다.

VIF 값이 높은 변수가 있다면 해당 변수를 제거하거나 모델링에서 제외하려 하였다.일반적으로 VIF 값이 1에 가까울수록 공선성이 거의 없음을 의미하며, 5 이상이면 공선성을 의심하고, 10 이상은 문제로 간주한다. 분석 결과, 모든 변수가 1에 가까운 VIF 값을 가졌음을 볼 수 있고, 이는 다중공선성이 거의 없는 데이터임을 시사한다.

**d. 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
```

로지스틱 회귀는 변수의 선형 관계를 평가하므로, 비선형적 영향이나 복잡한 상호작용을 간과할 수 있다. 따라서 선형적으로는 알 수없는 관계가 변수간에 존재하기 때문에 성능 저하가 일어난다고 판단했고 비선형 관계와 상호작용을 더 잘 포착할 수 있으며, 변수 중요도를 측정할 때 각 변수의 분할 기여도를 평가하는 RandomForest를 사용해 주요 변수를 선택하기로 하였다.

RandomForestClassifier를 사용해 랜덤 포레스트 모델을 생성하고 학습시켰다. fit 메서드로 모델을 학습한 후, predict로 예측 값을 얻는다. feature_importances 속성을 사용해 각 특성(변수)의 상대적 중요도를 확인할 수 있었다. 'oldpeak', 'thalach', 'ca', 'cp', 'thal', 'age', 'trestbps', 'exang', 'chol', 'slope', 'sex', 'restecg', 'fbs' 순으로 중요도를 가지고 있었고, Feature Importance 값의 누적 합을 계산하여, 80% 이상의 중요도를 설명하는 변수만 선택하여, 하위 5개의 컬럼은 제외하고 로지스틱 회귀를

재진행해보았다.

### 7. 로지스틱 회귀 분석II

기존의 로지스틱 회귀 분석과 동일하게 종속변수인 'target'데이터를 데이터셋과 분리해 X와 y로 구분하였고, 8:2의 비율로 test set과 train set을 랜덤샘플링을 통해 가졌다. StandardScaler를 사용하여 데이터를 표준화한 후, 로지스틱 회귀 분석 모델에 적용했다. 회귀 분석 결과 평가지표 값은 아래와 같았다.

```
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
```

Feature Selection을 적용하여 로지스틱 회귀 분석 모델의 성능을 개선하고자 하였다. 평가 결과, 특성 선택을 적용한 모델이 기존 모델에 비해 전반적으로 더 높은 분류 성능과 예측 정확도를 달성했음을 확인하였다.

#### a. 정확도(Accuracy) 개선

기존 모델은 전체 샘플 중 약 85.25%를 올바르게 분류하였으나, 특성 선택 이후 모델은 88.52%로 성능이 향상되었다. 이는 전체 샘플에서 올바른 예측 비율이 증가했음을 보여준다.

#### b. 정밀도(Precision)의 향상

기존 모델은 전체 샘플 중 약 85.25%를 올바르게 분류하였으나, 특성 선택 이후 모델은 88.52%로 성능이 향상되었다. 이는 전체 샘플에서 올바른 예측 비율이 증가했음을 보여준다.

#### c. 재현율(Recall)의 개선

클래스 0의 재현율은 86%로 유지되었으나, 클래스 1의 재현율이 84%에서 91%로 크게 증가하였다. 이는 모델이 실제 양성 데이터를 더 잘 식별할 수 있게 되었음을 나타내며, 잘못된 음성 예측(false negative)을 효과적으로 감소시켰다.

#### d. F1-Score 상승

클래스 0의 F1-Score는 85%에서 88%로, 클래스 1은 86%에서 89%로 상승하였다. 이는 Precision과 Recall 간의 균형이 기존 모델보다 더 잘 이루어졌음을 보여준다.

### 8. 결론

특성 선택 기법을 통해 모델의 주요 변수 'oldpeak', 'thalach', 'ca', 'cp', 'thal', 'age', 'trestbps', 'exang'만 활용함으로써, 로지스틱 회귀 모델의 성능이 전반적으로 개선되었음을 확인하였다. 이는 모델의 단순화를 통해 불필요한 변수를 제거하여 예측력을 강화한 결과로 해석할 수 있다. 특히, 클래스 1(양성 데이터)에 대한 재현율과 F1-Score가 크게 향상되었으며, 이는 해당 클래스의 예측 정확성을 높이는 데 기여하였다.

따라서 본 연구는 심혈관 질환 발병 여부 예측에서 RandomForest를 활용한 Feature Selection 기법이 로지스틱 회귀 모델의 성능 향상에 효과적임을 보여주었으며, 데이터 분석 및 모델 최적화 과정에서 중요한 도구로 활용될 수 있음을 시사한다. 향후 연구에서는 보다 다양한 특성 선택 기법 및 다른 머신러닝 모델과의 성능 비교를 통해 더욱 정교한 분석과 모델링이 이루어질 수 있을 것이다.

### 9. 기여점

본 연구는 로지스틱 회귀 분석을 활용하여 심장 질환의 주요 변수에 대한 명확한 통찰을 제공함으로써, 심장 질환 조기 진단과 예방, 공중보건 정책 개선에 기여할 수 있다. 또한, 효과적인 데이터 기반 의료 시스템 구축을 위한 기초 자료로 활용될 수 있을 것이다.

#### a. 심장 질환 주요 변수의 정량적 분석

로지스틱 회귀 분석을 통해 심장 질환 발생에 주요한 영향을 미치는 변수들을 도출하였다. 이는 기존의 정성적 또는 경험적 분석에 비해 보다 객관적이고 정량적인 결과를 제공하며, 심장 질환의 위험 요인을 명확히 이해하는 데 기여한다. 특히, 모델 성능 개선을 위해 특성 선택 기법을 도입하여 예측 모델의 단순화와 해석력을 동시에 강화하였다.

#### b. 조기 진단과 예방을 위한 실질적 활용 가능성

도출된 주요 변수는 심장 질환의 조기 진단과 고위험군 선별에 유용하게 활용될 수 있다. 의료진은 본 연구 결과를 참고하여 정기 검진에서 중요한 위험 요인을 모니터링하고, 고위험 환자에 대한 정밀 검사를 통해 심장 질환 예방 조치를 강화할 수 있다. 또한, 환자 개인에게는 생활습관 개선 및 건강 관리 방안을 제공하여 발병 가능성을 낮추는 데 도움을 줄 수 있다.

**c. 의료 자원 활용의 효율성 증대**

연구 결과를 통해 의료진은 심장 질환의 고위험군을 보다 효율적으로 선별할 수 있으며, 불필요한 검사 및 치료를 최소화할 수 있다. 그리고 조기 발견을 통한 치료 성공률 향상은 의료비 절감 및 의료 자원의 효율적 배분에 기여할 수 있다.

**d. 공중보건 정책 및 캠페인에의 기여**

본 연구에서 제시된 주요 변수는 심장 질환 예방 캠페인, 건강 교육 프로그램, 생활습관 개선 정책 등의 공중보건 정책 설계에 활용될 수 있다. 예를 들어, 도출된 위험 요인을 중심으로 심혈관 질환 고위험군 대상의 맞춤형 건강 관리 프로그램을 개발하거나, 예방 캠페인의 핵심 메시지로 사용 가능