2025 0619

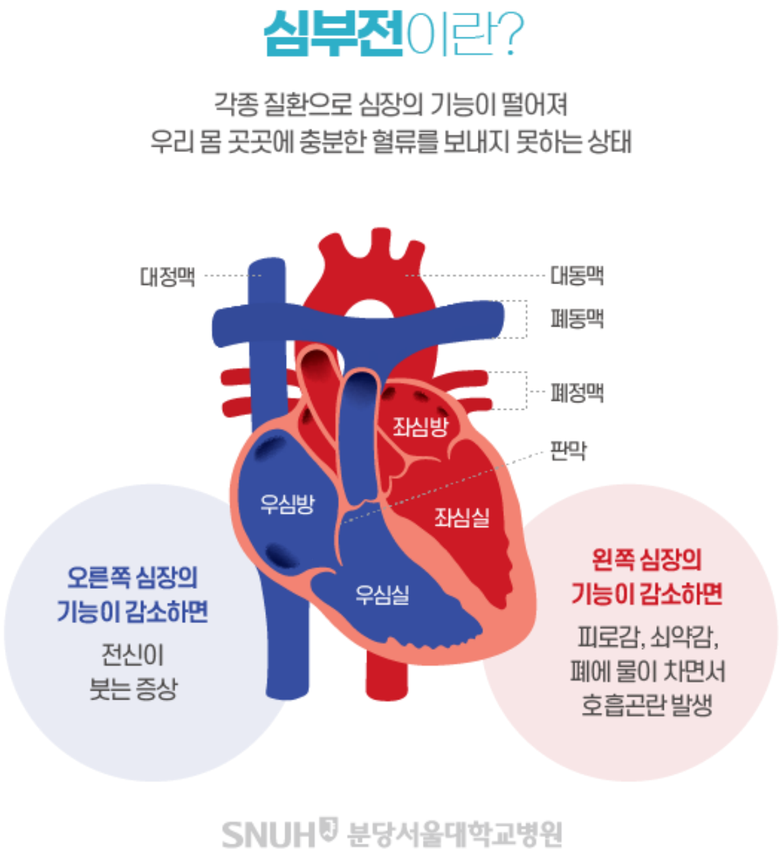
AI Advanced Programming

Heart Failure Prediction Dataset

정보통신공학과 2060025 손진승 (리더)
정보통신공학과 2060009 김재헌
정보통신공학과 2060021 박찬혁

프로젝트 목차

심부전(Heart Failure)이란?



心不全, **Heart Failure**

각종 심장질환으로 인해 심장에 구조적 혹은 기능적 이상이 생겨
심실의 혈액 충만 또는 박출에 이상이 발생한 상태

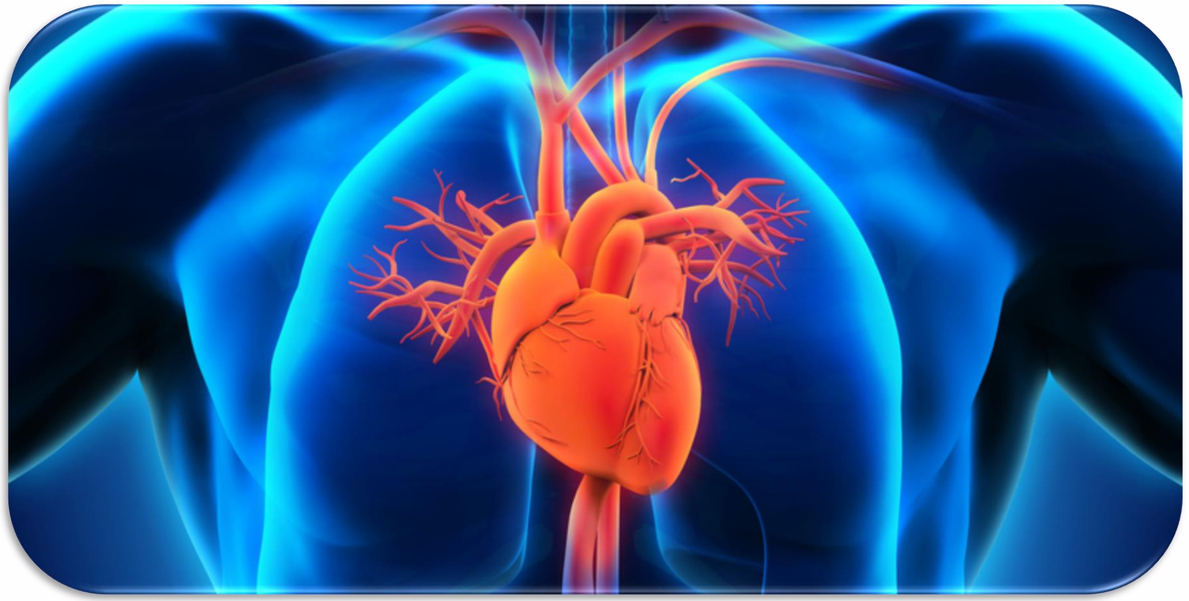
Ex) 피로, 호흡곤란, 부종 등

주제 소개

Kaggle → Heart Failure Prediction Dataset을 기반,
심부전 발생과 관련된 다양한 건강 지표 간의 연관성을 분석

심부전은 매우 중요한 건강 문제이며, 정기적인 검진과 예측
시스템을 통해 조기 발견과 체계적인 관리가 가능

다양한 건강 특성과 타겟 변수를 통해 심부전 발생과의
연관성을 분석할 수 있다는 점에서 이 주제를 선택



선정 이유

1. 의학적 중요성

- 심장질환은 한국 전체 사망 원인 중 2위
- 고령 인구 증가와 함께 심부전 환자 수 또한 지속적으로 증가
- 조기 발견 시 치료 효과가 높아 'AI 기반 예측 시스템의 필요성' 강조

2. AI 적용 적합성

- 심부전은 혈압,콜레스테롤,심박수 등 다양한 수치형 변수들과 연관
- 머신러닝 기반 분류 모델을 통해 예측 가능

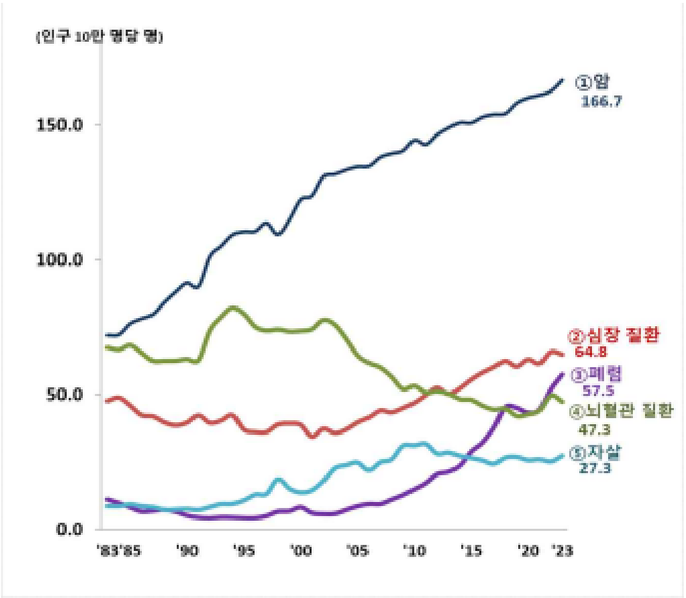
3대 사망원인은 암, 심장 질환, 폐렴 (전체 사망의 41.9%)

○ 10대 사망원인은 악성신생물(암), 심장 질환, 폐렴, 뇌혈관 질환, 고의적 자해(자살), 알츠하이머병, 당뇨병, 고혈압성 질환, 패혈증, 코로나19 순임.

<사망원인 순위 추이>

(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)

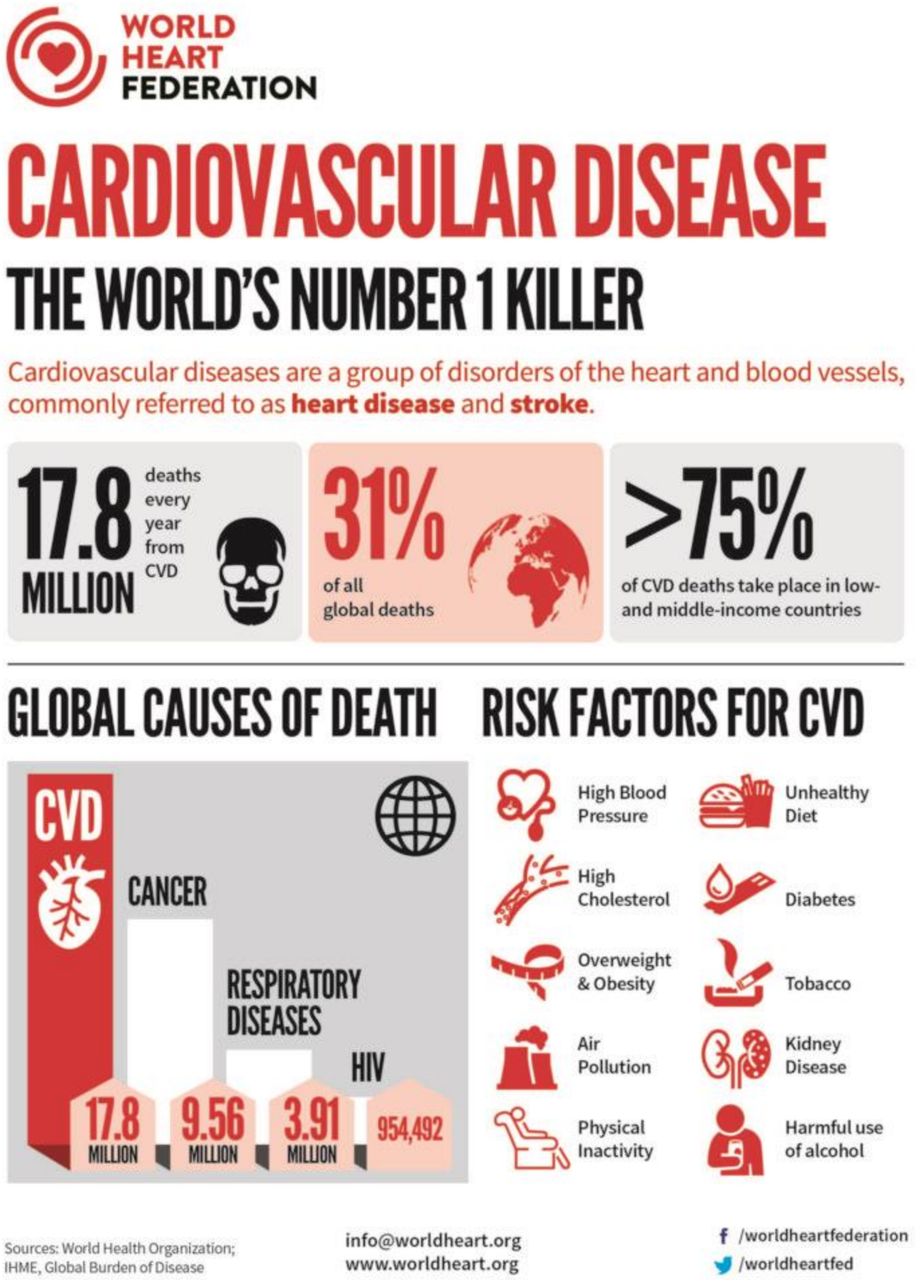


참고문헌 : https://kostat.go.kr/board.es?mid=a10301060200&bid=218

주제 설명

Heart Failure Prediction Dataset

- 심혈관 질환(CVDs)은 전 세계 사망 원인 1위로, 매년 약 1,790만 명의 생명을 앗아가며, 전체 사망의 약 31%를 차지함
- 심부전(Heart Failure)은 심혈관 질환(CVD)로 인해 발생하는 흔한 합병증으로 알려짐
- 이 데이터셋은 심부전 예측을 위해 혈압, 심전도, 가슴통증 등의 위험요소를 포함한 11개 특성을 수집함
- 위험요소가 하나 이상 있는 심혈관 질환 고위험군에 속하는 사람은 조기에 발견하여 관리해야 하며, 이때 머신러닝 모델이 큰 도움을 줄 수 있음



해결 목적

- 조기 발견의 어려움
심부전은 혈압, 심전도, 가슴통증 등 다양한 위험요소가 복합적으로 작용하여 발생하지만, 환자가 자각하기 어려운 경우가 많음
→ 데이터 기반 머신러닝 모델을 활용해 잠재적 고위험군을 조기에 식별 가능

- 의료진 진단 부담 증가
심혈관(부전) 환자는 다양한 지표를 고려해야 하므로 진단이 복잡함
→ 머신러닝 모델은 의료진의 진단을 지원하여, 보다 정확하고 빠른 결정을 도움

- 고령화 사회 대응 가능성
고령 인구 증가로 심부전 환자 수가 늘어날 것으로 예상됨
→ 조기 예측을 통해 고령화에 따른 의료 비용 부담 증가 문제에 간접적으로 대응 가능

Dataset

```python
import pandas as pd

data = pd.read_csv('heart.csv')

print("데이터 미리 보기:")
print(data.head())

print("₩n칼럼 목록:")
for col in data.columns:
    print("-", col)
```

```
데이터 미리 보기:
   Age  Sex  ChestPainType  ...  Oldpeak  ST_Slope  HeartDisease
0   40    0              0  ...      0.0         0             0
1   49    1              1  ...      1.0         1             1
2   37    0              0  ...      0.0         0             0
3   48    1              2  ...      1.5         1             1
4   54    0              1  ...      0.0         0             0

[5 rows x 12 columns]

칼럼 목록:
- Age
- Sex
- ChestPainType
- RestingBP
- Cholesterol
- FastingBS
- RestingECG
- MaxHR
- ExerciseAngina
- Oldpeak
- ST_Slope
- HeartDisease
>>>
```

변수명 (칼럼명)	의미 설명	데이터 형태
Age	환자의 나이 (만 나이 기준)	int
Sex	성별(M=0 / F=1)	Object -> int
ChestPainType	가슴 통증 유형 (ASY/NAP/ATA/TA) 0 ~ 3 (위험도 순)	Object -> int
RestingBP	안정 시 혈압 (mm Hg)	int
Cholesterol	혈중 콜레스테롤 수치 (mg/dL)	int
FastingBS	공복혈당 > 120mg/dL 여부 (0 : 아님 / 1 : 맞음)	이진형(0 or 1)
RestingECG	안정 시 심전도 결과 (Normal, LVH, ST) 0 ~ 2 (심각도 순)	Object -> int
MaxHR	최대 심박수(운동 중 기록된 값)	int
ExerciseAngina	운동 유발 협심증 여부 (Y / N)	Object
Oldpeak	ST 우울증 수치 (심장스트레스)	float
ST_Slope	ST 분절 기울기 (Up,Flat,Down) 0 ~ 2 (위험도 순)	Object -> int
HeartDisease	타겟 변수 (심부전 유무) (0 : 없음 / 1 : 있음)	이진형(0 or 1)

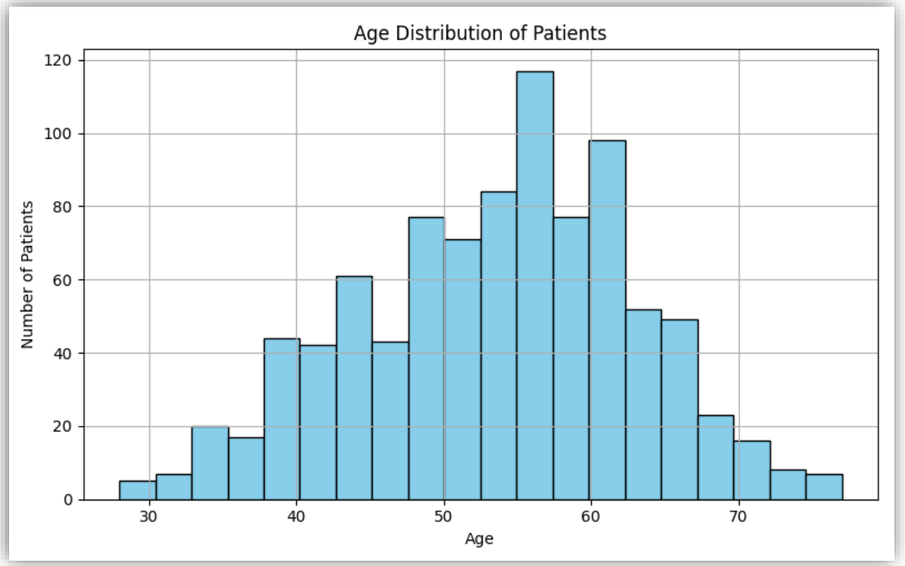
Dataset

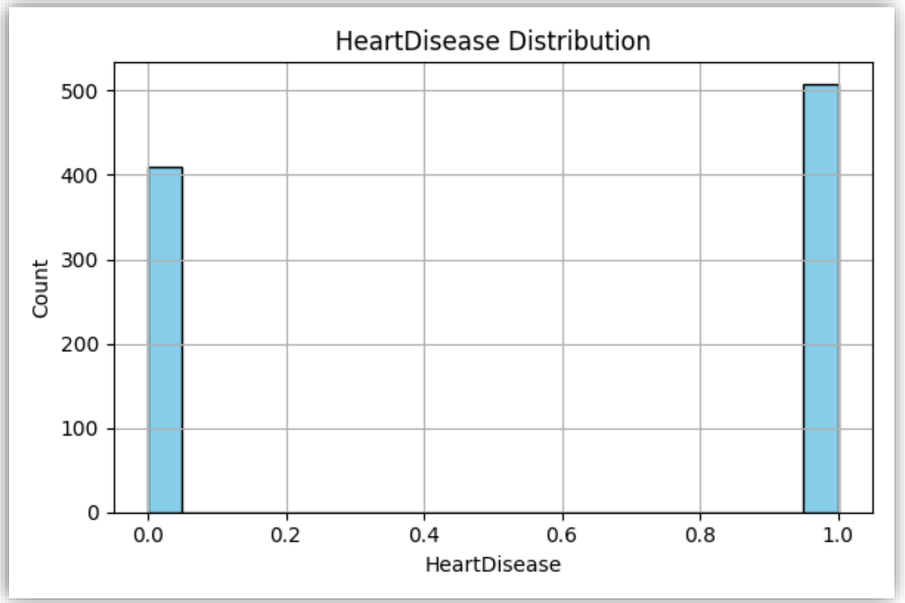
◆ Dataset
- 총 918명 환자 데이터
- 11개 입력 특성과 1개 타겟 변수 (HeartDisease)
- 결측치 없음 (전처리 용이)

◆ 데이터 특징 요약
- 주요 수치형 변수: 나이, 혈압, 콜레스테롤 등
- 일부 범주형 변수는 수치형으로 인코딩 처리 : 성별(M/F→0/1), 가슴 통증 유형 등
- 심부전 발생 여부 타겟 (1/0) 분포 약 55:45

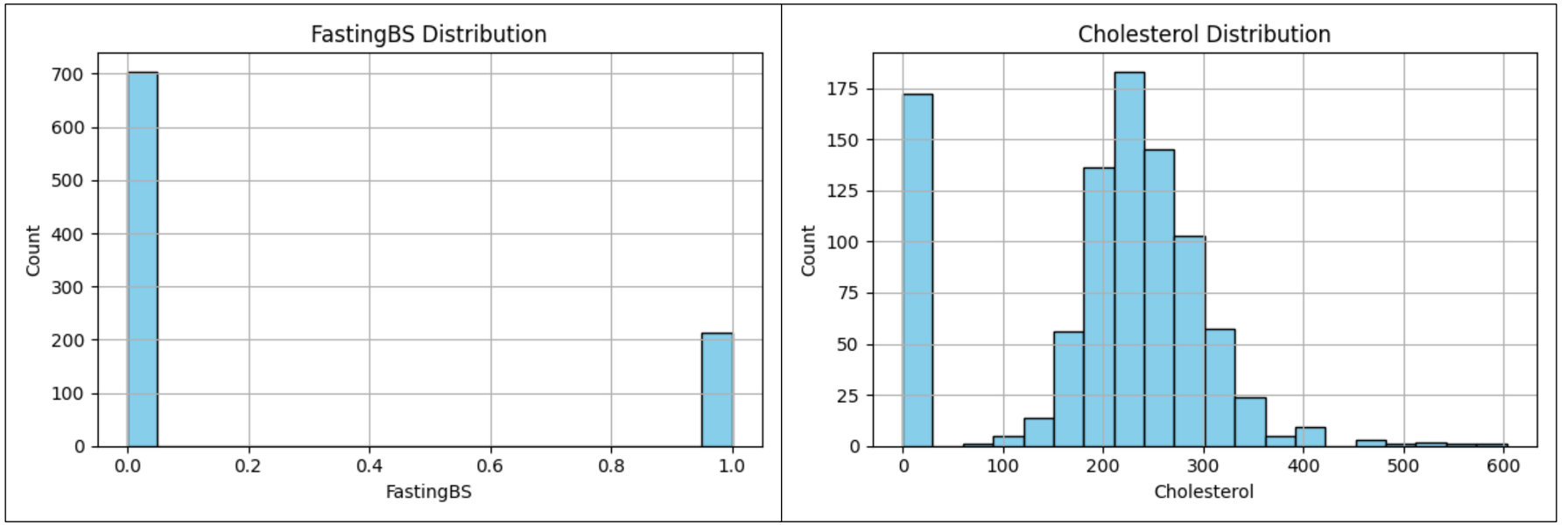
◆ 추가 분석 (Age)
- 환자 다수는 50대 중후반과 60대 초반에 분포
- 고령 환자 비율이 높음 ➔ 심부전과 연관

* 대표적인 수치형 변수인 '나이(Age)'에 대한 분포를 오른쪽 히스토그램으로 표시

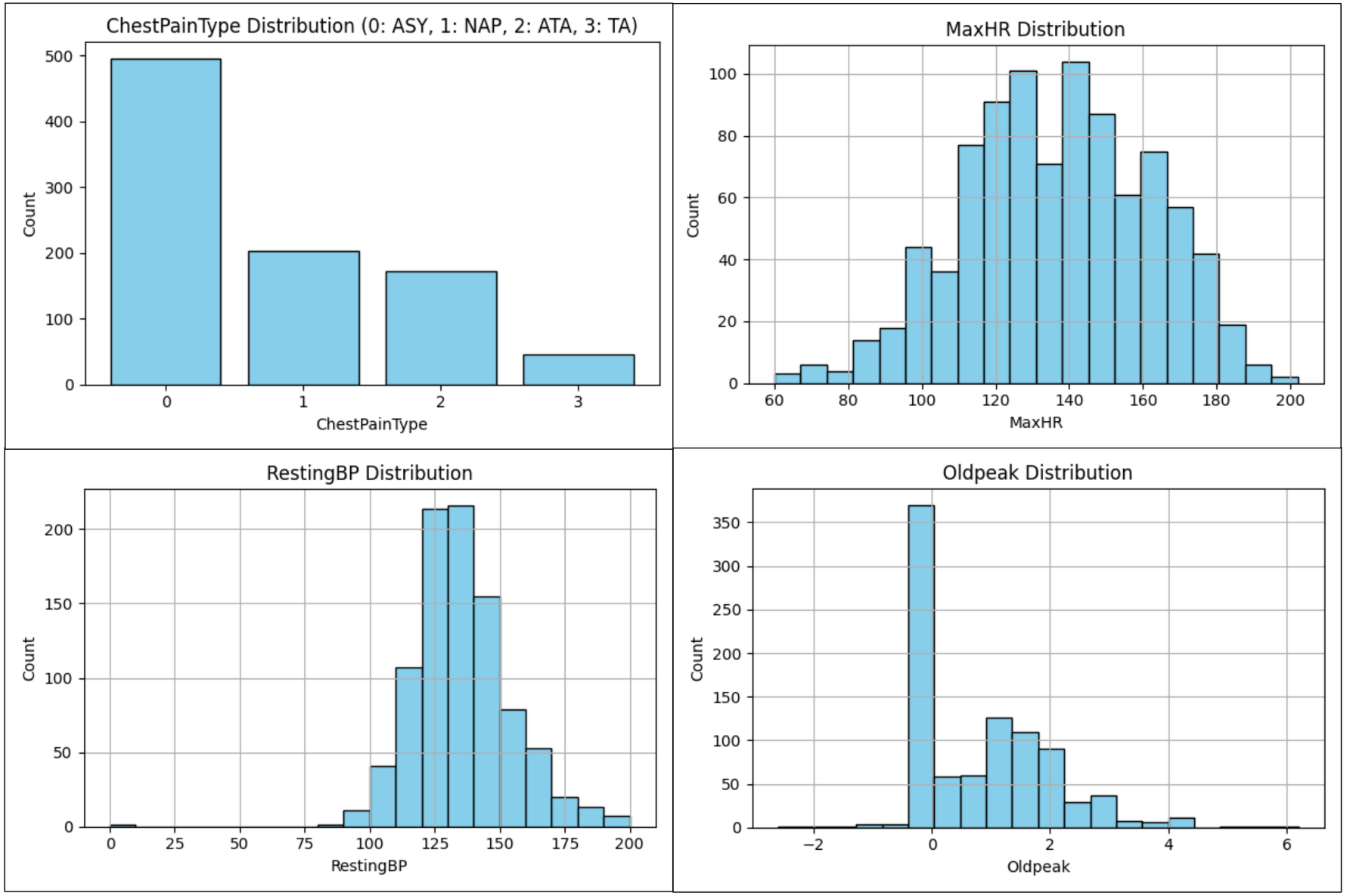




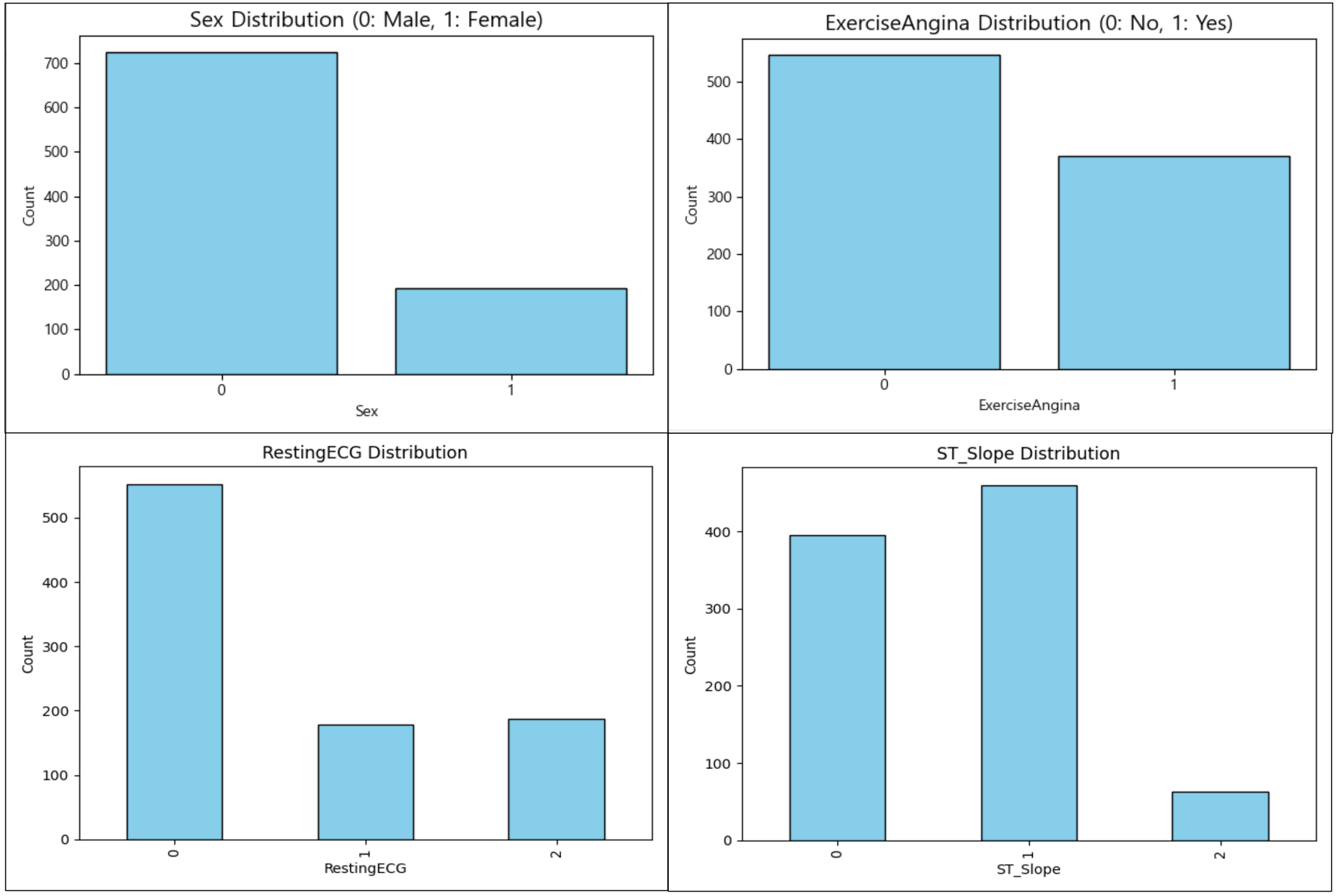
Histogram


FastingBS Distribution
Count
FastingBS
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
600
700
Cholesterol Distribution
Count
Cholesterol
0
100
200
300
400
500
600
0
25
50
75
100
125
150
175

Histogram


ChestPainType Distribution (0: ASY, 1: NAP, 2: ATA, 3: TA)
Count
ChestPainType
0
1
2
3
MaxHR Distribution
Count
MaxHR
RestingBP Distribution
Count
RestingBP
Oldpeak Distribution
Count
Oldpeak

Histogram


Sex Distribution (0: Male, 1: Female)
Count
Sex
0
1
ExerciseAngina Distribution (0: No, 1: Yes)
Count
ExerciseAngina
0
1
RestingECG Distribution
Count
RestingECG
0
1
2
ST_Slope Distribution
Count
ST_Slope
0
1
2

Neural Network Architecture

입력 변수 (X) → 총 11개 특성
- Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope

출력 변수 (y) → HeartDisease (0: 심부전 X / 1: 심부전 O)

신경망 구조
- 입력층(11개 노드) → 은닉층 1개 → 출력층(Softmax, 노드 1개)

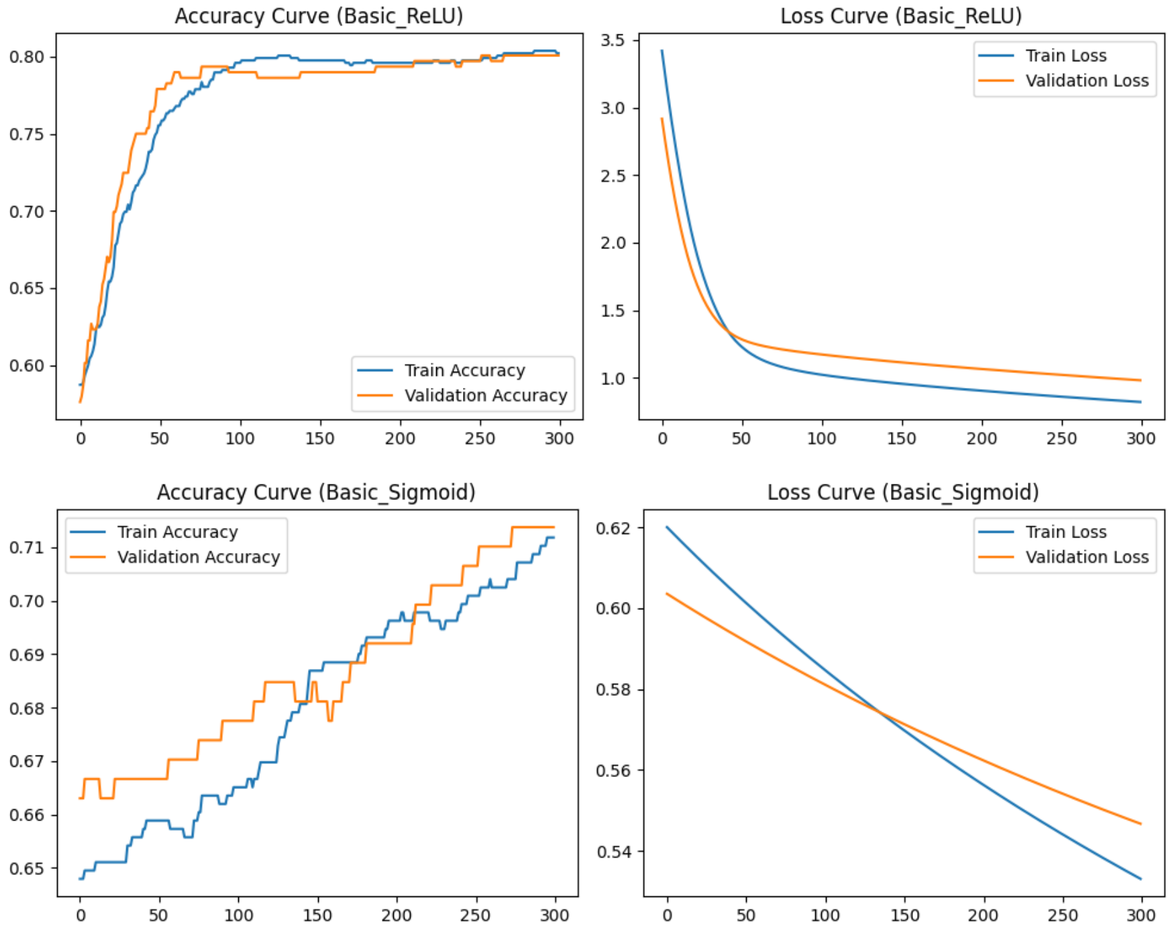
학습 목적
- 신경망이 심부전 발생 여부를 조기 예측할 수 있도록 학습하는 것

비율 설정
- train = 70%, test = 30%

ReLU, Sigmoid

1. 실험 목적
- 수업 시간에 배운 오차역전파 구조를 기반으로 ReLU와 Sigmoid 활성화 함수의 성능을 비교

2. 실험 조건
① 기본 실험:
- 히든층 노드 수: 8개
- 학습 반복 횟수(Epoch): 300회

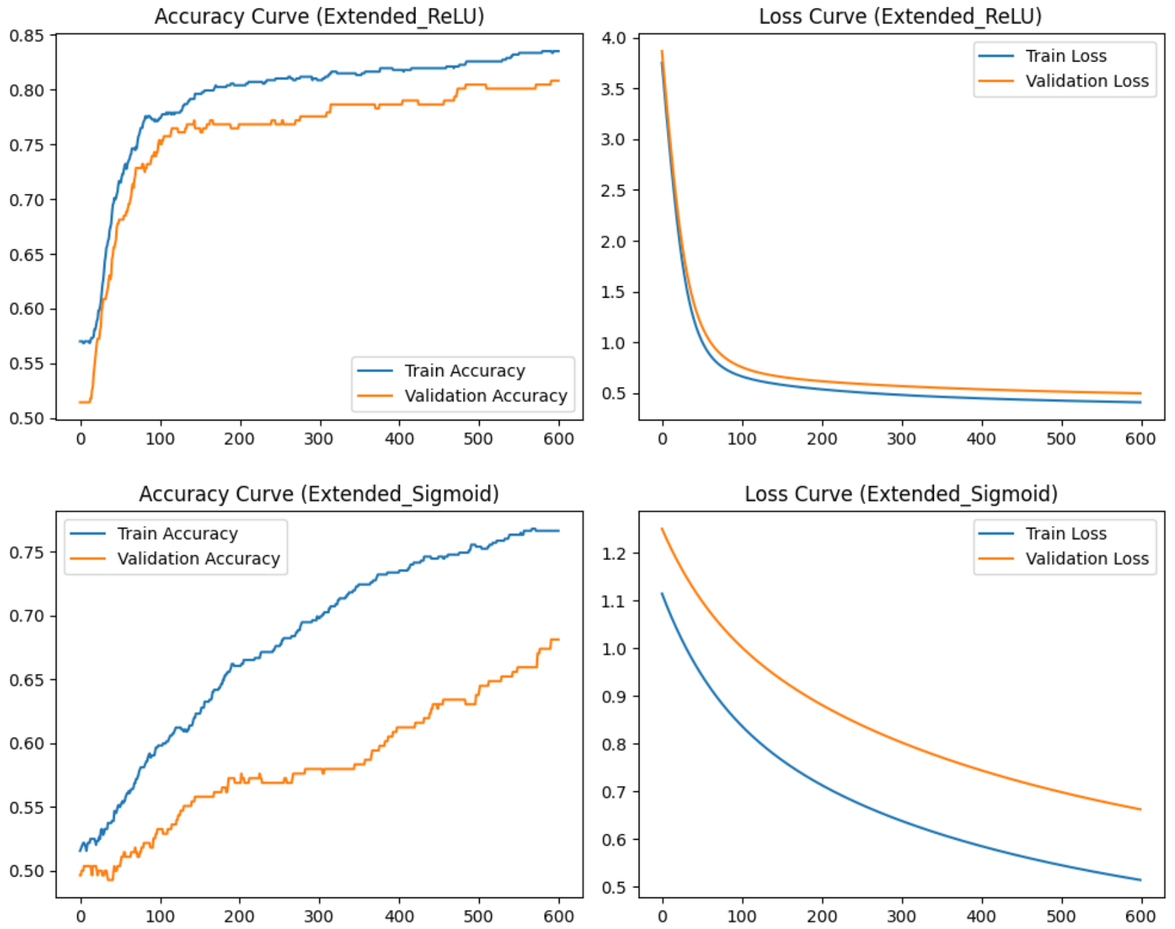


ReLU, Sigmoid

② 확장 실험:
- 히든층 노드 수: 16개
- 학습 반복 횟수(Epoch): 600회

3. 비교 목적
- 두 활성화 함수의 학습 경향 및 성능 차이를 동일한 구조 내에서 조건만 변경하여 확인

4. 시각화 내용
- 오른쪽 그래프는 각 실험 조건에서의 Cross Entropy 손실 값(Loss)의 변화 추이 (Loss Curve)



ReLU, Sigmoid

2. 결과 비교 요약

비교 항목	기본구조 ReLU	기본구조 Sigmoid	확장구조 ReLU	확장구조 Sigmoid
Train Accuracy	80.22%	71.18%	83.49%	76.64%
Validation Accuracy	80.07%	71.38%	80.80%	68.12%
Train Loss	0.8215	0.5331	0.4099	0.5141
수렴 안정성	수렴 양호	수렴 느림	수렴 양호	수렴 불안정
과적합 여부	비교적 양호	안정적	약간 과적합	과적합 발생
일반화 성능	우수	낮음	우수	제한적

- 전반적으로 ReLU가 Sigmoid보다 전반적 성능이 우수

- 특히 확장 실험에서도 ReLU가 안정적인 수렴 및 높은 정확도 유지

- Sigmoid는 특히 Hidden Size가 커지면 과적합 경향이 더 강해짐 (은닉층 기준)

→ 따라서 본 프로젝트에서는 기본 구조 + ReLU 조합이 가장 적합한 모델 구성으로 판단됨

```
===== Basic_ReLU =====
최종 Train Accuracy: 80.22%
최종 Validation Accuracy: 80.07%
최종 Train Loss: 0.8215
최종 Validation Loss: 0.9820

===== Basic_Sigmoid =====
최종 Train Accuracy: 71.18%
최종 Validation Accuracy: 71.38%
최종 Train Loss: 0.5331
최종 Validation Loss: 0.5467

===== Extended_ReLU =====
최종 Train Accuracy: 83.49%
최종 Validation Accuracy: 80.80%
최종 Train Loss: 0.4099
최종 Validation Loss: 0.4987

===== Extended_Sigmoid =====
최종 Train Accuracy: 76.64%
최종 Validation Accuracy: 68.12%
최종 Train Loss: 0.5141
최종 Validation Loss: 0.6623
```