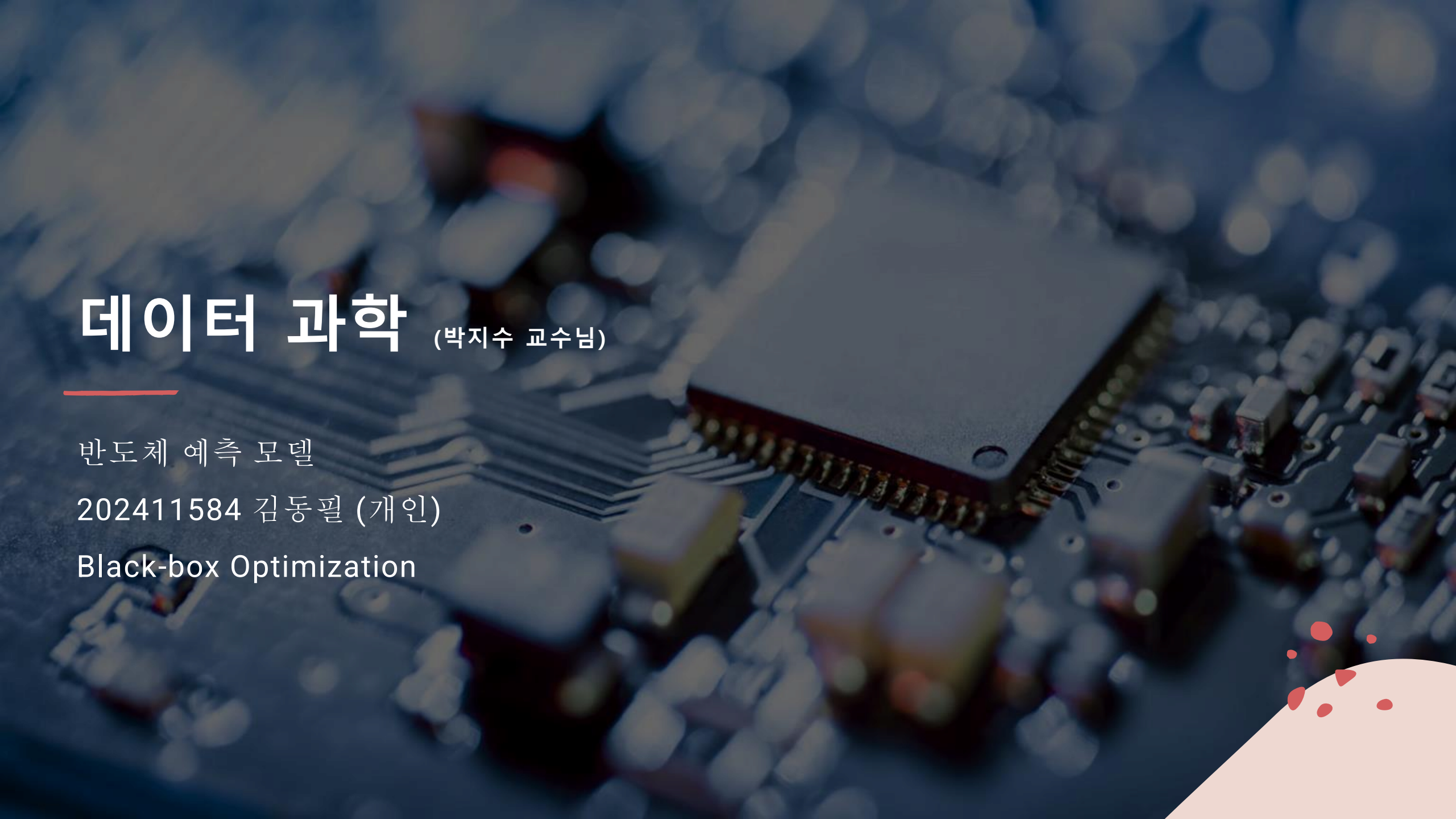


데이터 과학 (박지수 교수님)

반도체 예측 모델

202411584 김동필 (개인)

Black-box Optimization

목적.

최적의 모델을 개발하기 위해, 오프라인 모델 기반 최적화 기법을 활용하여 데이터 분포와 최적화된 파라미터의 균형점을 잘 찾아야 합니다.

이를 통해 black box 문제에 대한 AI 알고리즘의 성능을 최대한 향상시키는 것이 목표입니다.

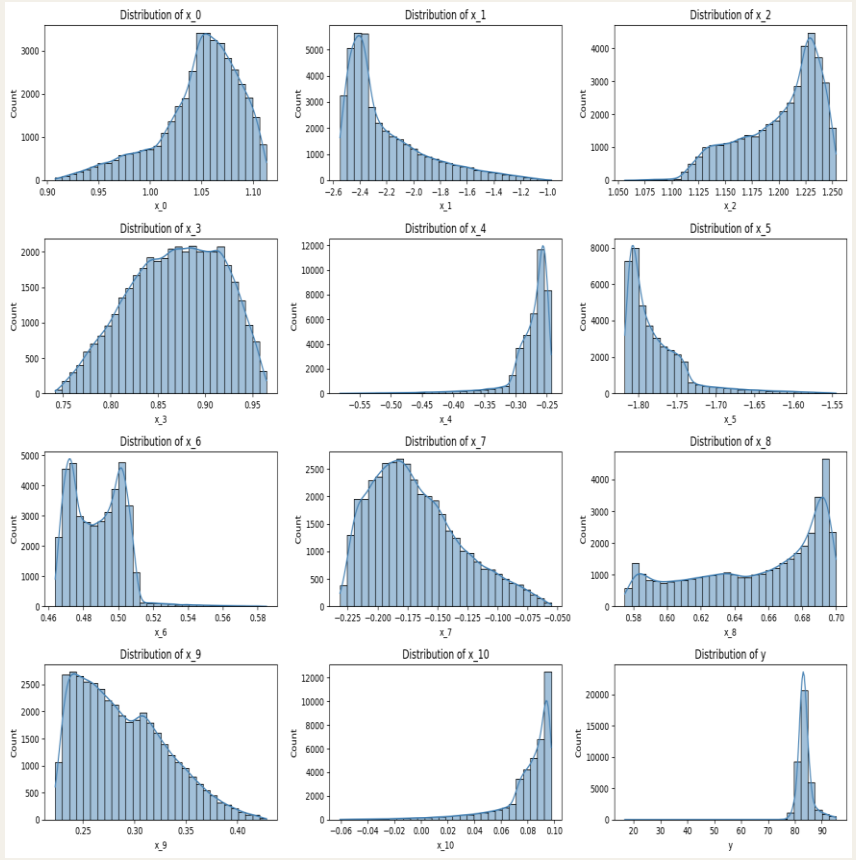
주어진 입력 변수 x_0부터 x_10까지의 값을 통해 예측된 타겟 변수 y의 값 중 상위 10%를 찾아내고 이 예측된 상위 10%의 데이터 중 상위 데이터가 얼마나 포함되어 있는지를 측정하고 평가합니다.

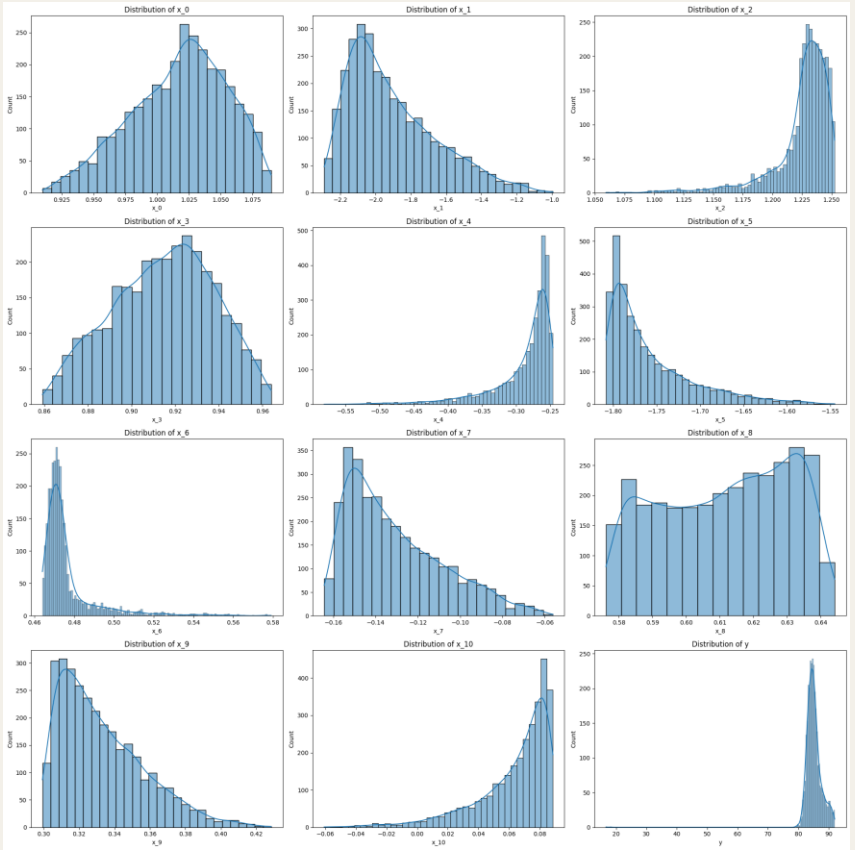
프로젝트의 전체적인 흐름

1.라이브러리 임포트:
데이터 조작, 시각화, 전처리, 모델링에 필요한 다양한 라이브러리를 불러옵니다.
2.EDA 함수 정의:
데이터의 기본적인 특성을 파악하고 시각화하기 위한 함수를 정의합니다.
3.데이터 로드 & EDA 수행:
학습 데이터와 테스트 데이터를 불러와 EDA를 수행하여 데이터의 특성을 파악합니다.
4.피처/타겟 분리:
학습 데이터에서 피처(X)와 타겟(y)을 분리하고, 테스트 데이터에서 피처만을 추출합니다.
5.결측치 처리 + 피처 전처리:
고급 결측치 처리를 통해 데이터를 정제하고, 상호작용 피처를 추가하며, 이상치를 클리핑하여 데이터의 품질을 향상시킵니다.
6.스케일링 + 데이터 분할:
데이터를 표준화하고, 학습 세트와 검증 세트로 분리하여 모델 학습과 평가를 준비합니다.
7.커스텀 리콜 메트릭 정의:
특정 성능 지표를 측정하기 위해 커스텀 리콜 메트릭을 정의합니다.
8.랜덤 포레스트 모델 학습 및 튜닝:
랜덤 포레스트 모델을 학습하고, 그리드 서치를 통해 최적의 하이퍼파라미터를 찾습니다.
9.XGBoost 모델 학습 및 튜닝:
XGBoost 모델을 학습하고, 랜덤 서치를 통해 최적의 하이퍼파라미터를 찾습니다.
10.LightGBM 모델 학습 및 튜닝:
LightGBM 모델을 학습하고, 랜덤 서치를 통해 최적의 하이퍼파라미터를 찾습니다.
11.딥러닝 모델 학습:
심층 신경망(DNN)을 구축하고, 배치 정규화와 드롭아웃을 통해 과적합을 방지하며, 조기 중단을 설정하여 효율적으로 학습을 진행합니다.
12.앙상블:
여러 모델의 예측 결과를 가중 평균하여 최종 예측값을 생성합니다.
13.테스트 데이터 예측 & 제출:
학습된 모델들을 사용하여 테스트 데이터에 대한 예측을 수행하고, 이를 제출 형식에 맞게 저장합니다.

분포 히스토그램

왼쪽(본인), 오른쪽(대회)





Box plot

(1) x_0, x_1, x_2, x_3, x_7, x_9, x_8

•분포:

- 이 변수들은 비교적 **정상적인 분포**를 가지고 있으며, 중앙값이 박스의 중간에 위치.
- 이상치가 거의 없거나 발견되지 않음.

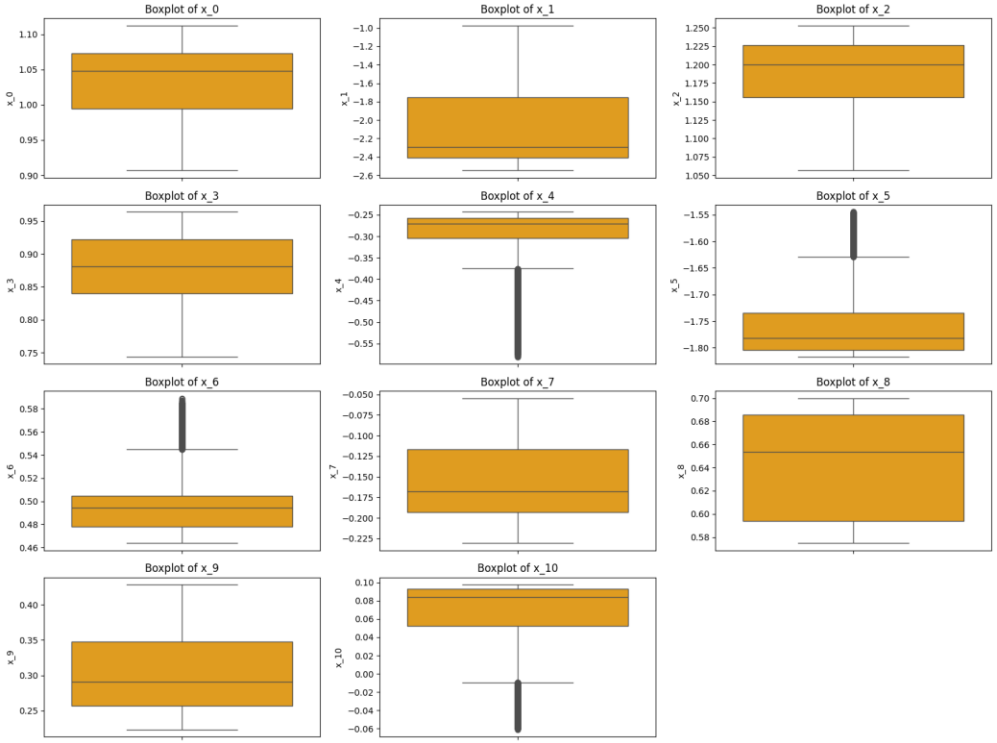
•해석:

- 해당 변수들은 데이터의 변동성이 적으며, 대부분의 값이 IQR 범위 안에 포함됨.
- 학습 모델에 안정적으로 기여할 가능성이 높음.

(2) x_4

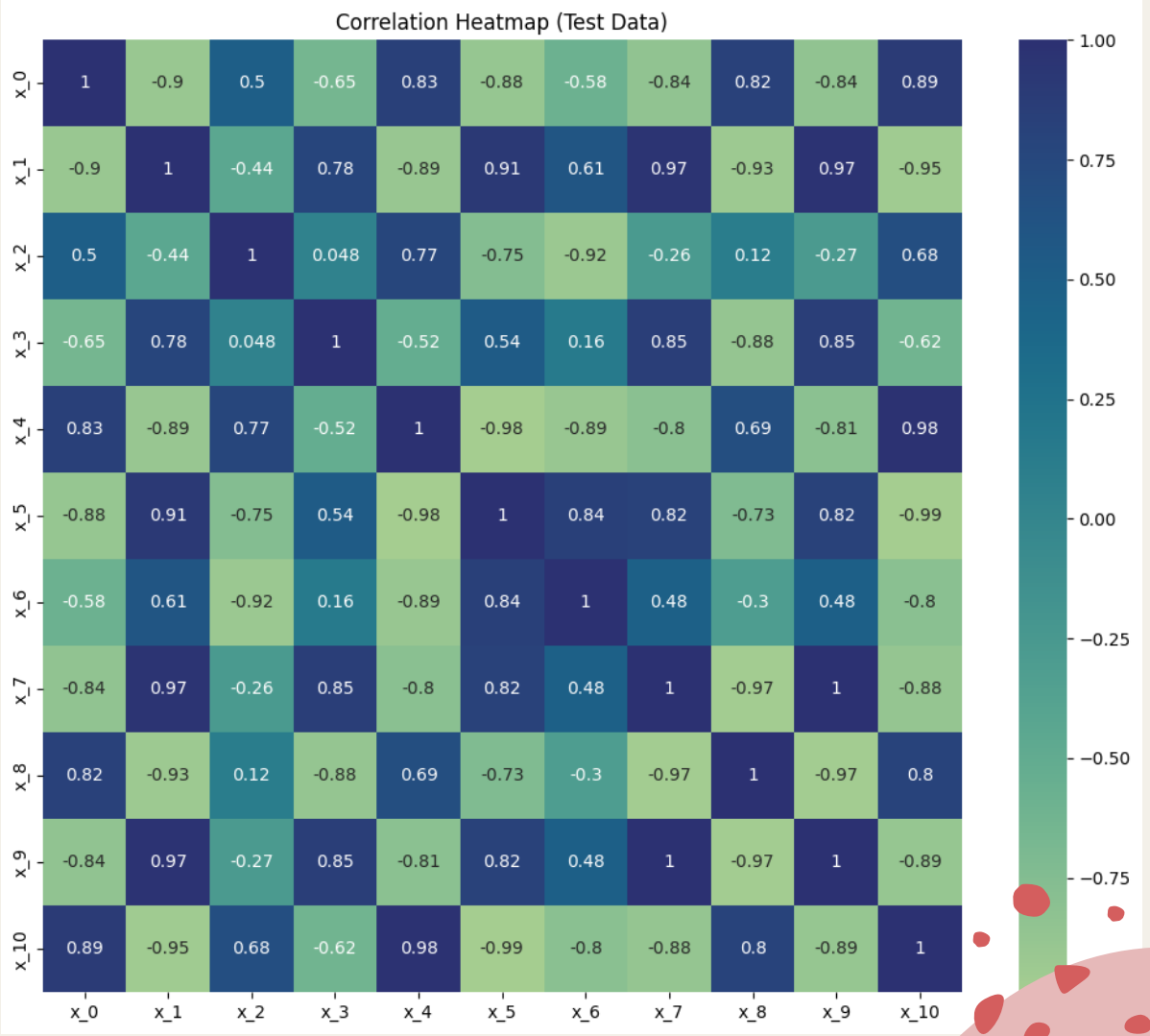
•분포:

- **중앙값이 박스 하단에 가까움**, 데이터가 약간 왼쪽으로 치우친(Skewed) 분포를 가질 가능성이 있음.
- Whisker 하단에 다수의 **이상치**가 존재.



Heat map

- **상관계수의 의미**:
- **양의 상관관계 (+)**: 한 변수가 증가하면 다른 변수도 증가하는 경향이 있음.
- **음의 상관관계 (-)**: 한 변수가 증가하면 다른 변수는 감소하는 경향이 있음.
- **상관관계 없음 (0에 가까움)**: 두 변수 간에 선형적인 관계가 거의 없음.
- **상관계수의 강도**:
- **0.7 이상 또는 -0.7 이하**: 강한 상관관계.
- **0.3 ~ 0.7 또는 -0.3 ~ -0.7**: 중간 정도의 상관관계.
- **0.3 미만 또는 -0.3 초과**: 약한 상관관계.



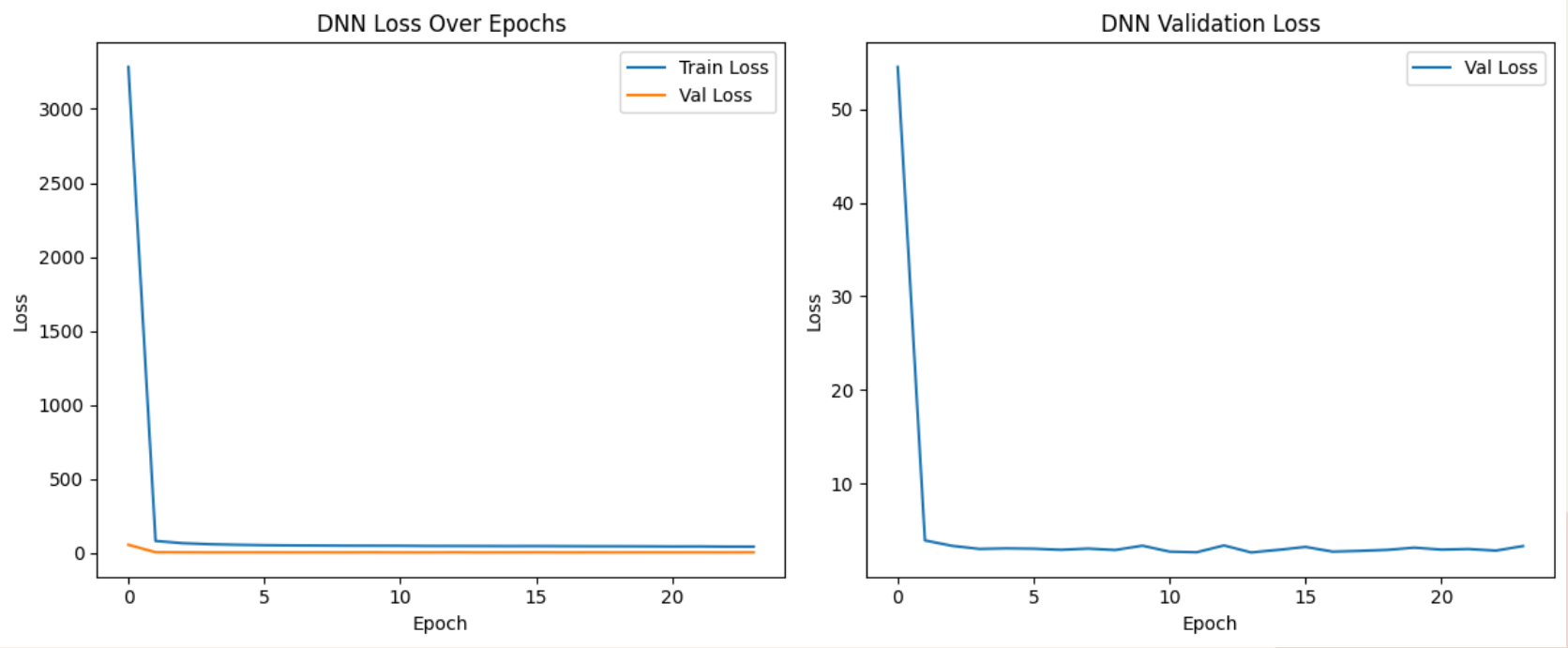
DNN Loss 손실값

상위 10% 예측 recall 값이 0.8905임을 확인할 수 있습니다.

모델 성능은 89~90%의 정확도를 가집니다.

251/251 ——————————————————————— **1s** 2ms/step [Val] DNN Recall: 0.8905 Model Weights based on Recall: RF=0.25, XGB=0.25, LGBM=0.25, DNN=0.25 [Val] Ensemble Recall: 0.8955 **156/156** ——————————————————————— **0s** 2ms/step

---



깃 허브

내용이 ppt에 담기에는 지나치게 많아 추가 설명은 깃허브에서 하겠습니다..
죄송합니다

RlEHVL/semiconductor-research