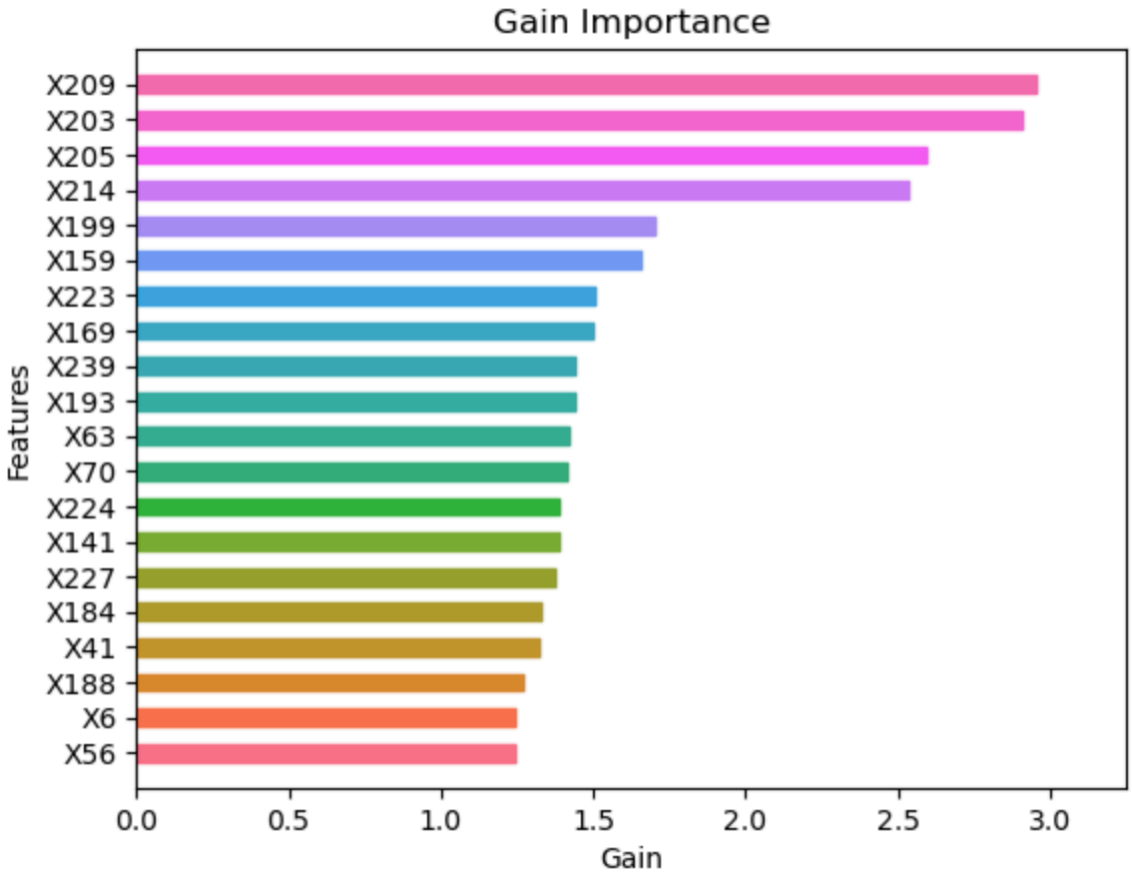
[데이터셋 설명]

다음 데이터셋(Dataset.csv)을 통해 제품의 불량 유형을 분류하는 인공지능 모델을 구축하고자 한다. 입력변수는 반도체 공정에서 모니터링된 수치형 변수들(X1~X240)이며, 제품 불량의 유형은 4가지이다.

[문제 1] XGBoost 알고리즘을 활용한 모델 구축 및 평가(5점)

XGBoost(XGBClassifier)를 활용하여 모델을 구축하고, **아래의 표**를 작성하시오. 또한, 학습된 XGBClassifier를 통해 'Gain'을 기준으로 상위 20개 변수의 **변수중요도**를 추출하고, 그래프를 출력하시오. (학습: 검증: 테스트 비율은 5:3:2로 하고, Z-score 기준 변수 스케일링을 진행한 후, 실습시간에 배운 XGBClassifier의 적절한 하이퍼파라미터들을 활용하여 모델을 구축하시오.)

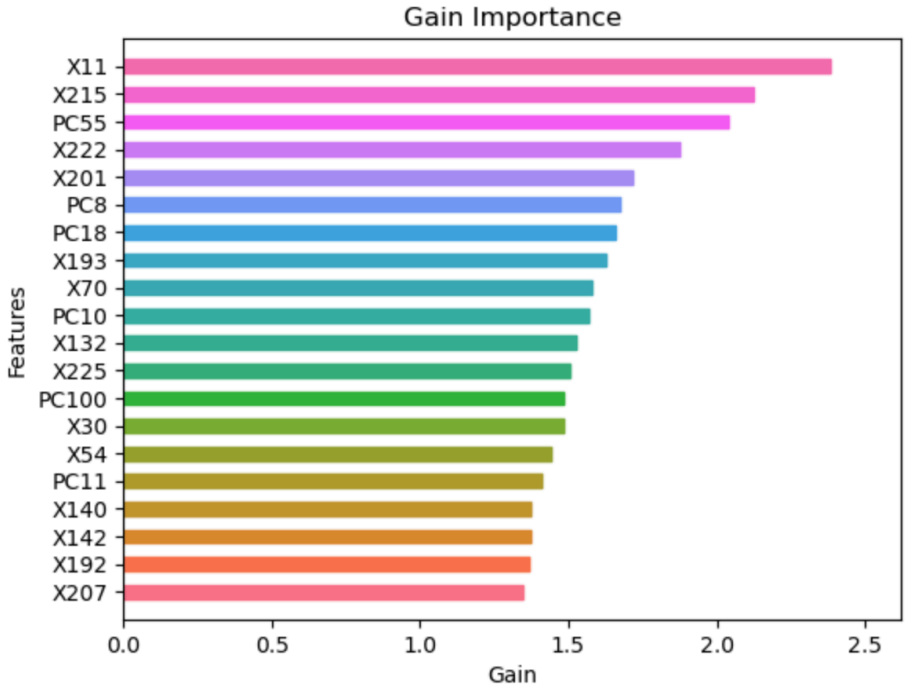
	F1-score	Accuracy
학습 데이터셋	1	1
검증 데이터셋	0.6499	0.6499
테스트 데이터셋	0.6685	0.6682



[문제 2] PCA알고리즘을 통한 차원 축소 및 변수 활용(5점)

학습데이터에 PCA알고리즘을 적용하여 분산설명을 80%이상 할 수 있는 만큼의 주성분을 데이터셋에 입력변수로 추가하고, 위와 동일한 방법으로 모델을 구축 및 분석하시오. 학습된 XGBClassifier를 통해 'Gain'을 기준으로 상위 20개 변수의 **변수중요도**를 추출하고, 위에서 선택된 변수와의 차이(PCA를 통한 차원 축소 변수의 영향력)를 분석하시오.

	F1-score	Accuracy
학습 데이터셋	1	1
검증 데이터셋	0.6533	0.6530
테스트 데이터셋	0.6649	0.6648




```
X209이 가장 크게 작용한 주성분: PC76
해당 주성분에서 X209의 기여도: -0.2142671634011294
```

위에서 구한 XGBoost 에서 Gain importance 가 제일 큰 변수가 PC1(T1) 에 속하지 않는다.

그 이유는 PCA 가 비지도학습(y값을 학습하지 않는)을 하고, PC1(T1) 에는 Z-Score 변환하고나서 분산이 가장 큰 변수가 선택되기 때문이다.

[문제 3] Blending을 통한 모델 성능 개선(5점)

[문제 2]에서 구축한 데이터셋(원본 데이터셋+주성분 데이터셋)에 추가적으로 Blending 기법을 통해 성능 개선을 하고자 한다. 이때, Blending의 Base모델로 활용할 모델은 Logistic Regression, LightGBM, SVM, RandomForest이며 메타모델은 XGBoost로 설정하여 위의 문제들과 성능 비교를 진행한다.

	F1-score	Accuracy
학습 데이터셋	-	-
검증 데이터셋	1	1
테스트 데이터셋	0.7046	0.7047

메타 모델은 Valid Data set으로 만들기 때문에 Train Data set에 대한 성능 점수는 없다.

각 모델에서 돌린 y값을 다시 학습 Data로 사용하게 되니까 성능이 많이 개선되는 것을 확인하였다.

	X1	X2	…	X240
Train				
Valid				
Test				

[문제1] 데이터셋 형태

	X1	X2	…	X240	PC_1	…	PC_k
Train							
Valid							
Test							

[문제2] 데이터셋 형태

	X1	X2	…	X240	PC_1	…	PC_k				
Train											
								X_LR	X_LGBM	X_SVM	X_RF
Valid											
Test											

[문제3] 데이터셋 형태

- 모든 문제는 모델의 성능과 상관없이 ① 정확한 분석 절차(코드 포함)와 ②분석 결과에 대한 해석을 바탕으로 평가함
- 제출파일은 분석 보고서(pdf 파일)와 코드(ipynb 파일)로 한다.