제2회 K-인공지능 제조데이터 분석 경진대회

제조공정 자원 최적화를 위한 불량 예측시스템의 오늘과 내일

(부제) SMOTE·불량 위험군 파생변수를 활용한 불량 예측시스템 및 수도레이블링을 통한 고도화방안 제안

(전자제조 9팀) 황선진, 이힘찬

목차

1. 문제정의

공정 개요

- ✓ 용해공정 : 분말 원재료를 액상 원재료에 녹이는 공정
- ✓ 본 용해공정은 분말 유크림, 기능성 조제 분말 등을 생산하는 식품제조업의 용해공정
  SD/MSD 건조생산라인의 원료 전처리 작업의 첫 번째 단계
- ✓ 용해공정의 품질은 용질과 용매의 화학적 특성, 용질과 용매의 상대적 용향, 용해온도, 물리적 힘 등에 영향을 받음

이슈 사항

- ✓ 용해품질에 영향을 미치는 다양한 요인들이 항상 존재
- ✓ 용해품질을 유지를 위해 상황에 따른 적절한 운영값 설정 필요, 그러나 일반적으로 현장작업자의 경험과 노하우에 의존해 대처
- ✓ 인력 공백이 생기는 경우 대처가 어렵고, 품질 이상 수준이 높아지면 자원 손실까지 이어짐

1. 문제정의

해결 방안

- 데이터를 기반으로 품질 이상을 미리 예측하는 **불량 예측시스템을 구축**하고,
  불량이 예상될 경우, 작업자에게 알리고 작업자는 적절한 설비운영값을 설정하여 최종 품질불량을 예방

- 또한, **실시간으로 얻어지는 데이터를 활용하는 프로세스를 마련**해 불량 예측시스템을 더 강건하게 만듦

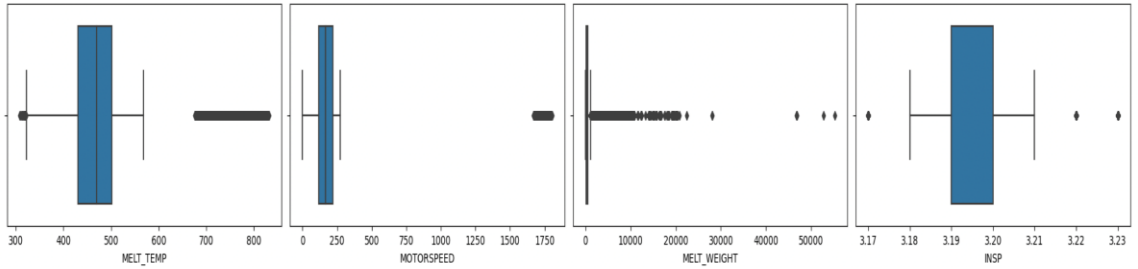
"SMOTE·불량 위험군 파생변수를 활용한 불량 예측시스템 및 수도레이블링을 통한 고도화방안 제안"

2. 데이터 처리 과정

- 데이터 확인 및 EDA

독립변수

- ✓ MELT_TEMP(온도), MOTORSPEED(모터속도), MELT_WEIGHT(중량), INSP(수분함유량) : 총 4개의 정수형 데이터(날짜 및 인덱스 제외)
- ✓ 데이터 개수 : 835,200개
- ✓ Boxplot 시각화 : 특별한 이상치 없는 것으로 판단

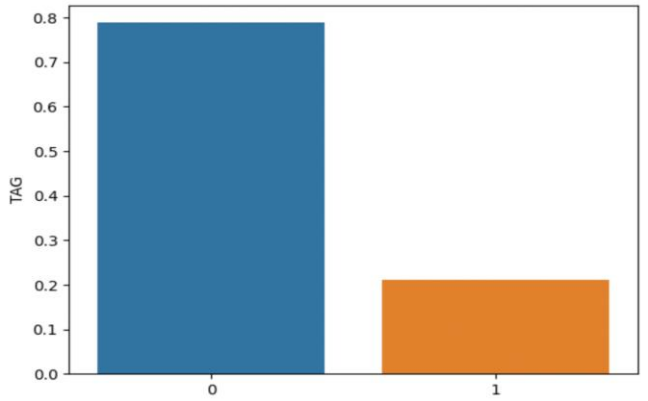


- ✓ 상관분석 확인(종속변수 포함) : MELT_TEMP와 종속변수간 상관관계가 가장 높음

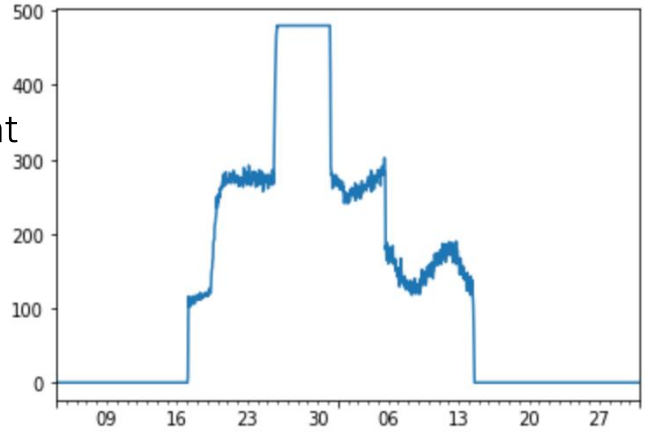
	MELT_TEMP	MOTORSPEED	MELT_WEIGHT	INSP	TAG
MELT_TEMP	1.000000	0.944929	-0.000336	0.916295	-0.310586
MOTORSPEED	0.944929	1.000000	0.000123	0.887813	-0.264693
MELT_WEIGHT	-0.000336	0.000123	1.000000	-0.000005	0.012084
INSP	0.916295	0.887813	-0.000005	1.000000	-0.272580
TAG	-0.310586	-0.264693	0.012084	-0.272580	1.000000

종속변수 (0:OK / 1:NG)

- ✓ 종속변수분포
  : 8대 2 수준의 불균형분포



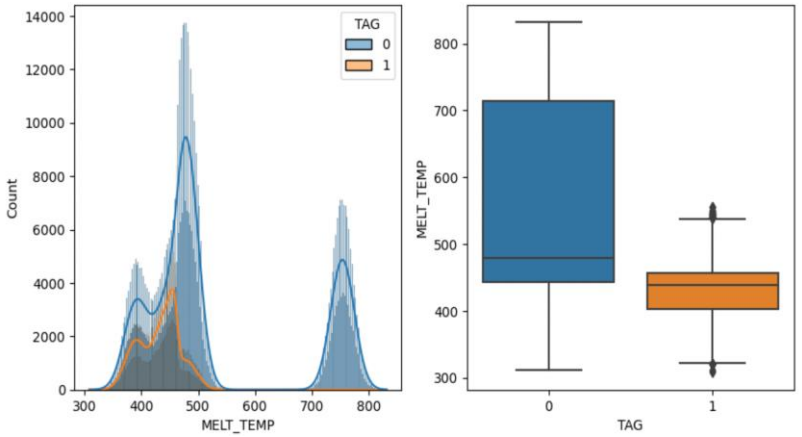
- ✓ 종속변수 시계열(시간별) count
  : 특정 구간에 모인 불균형분포



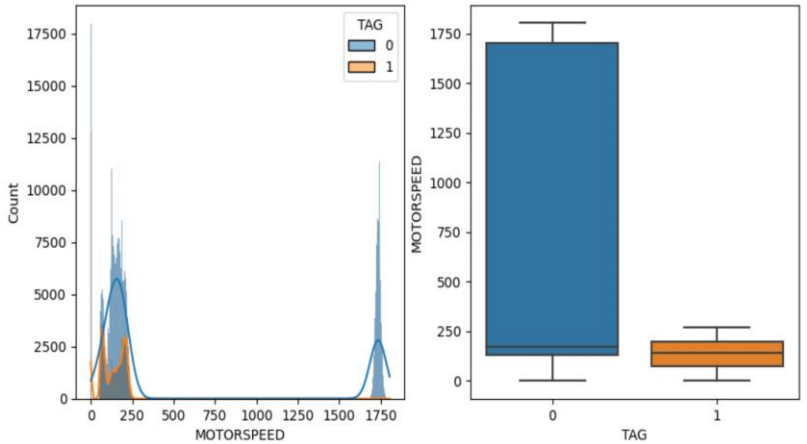
2. 데이터 처리 과정

- OK(0)·NG(1)별 독립변수 패턴 EDA

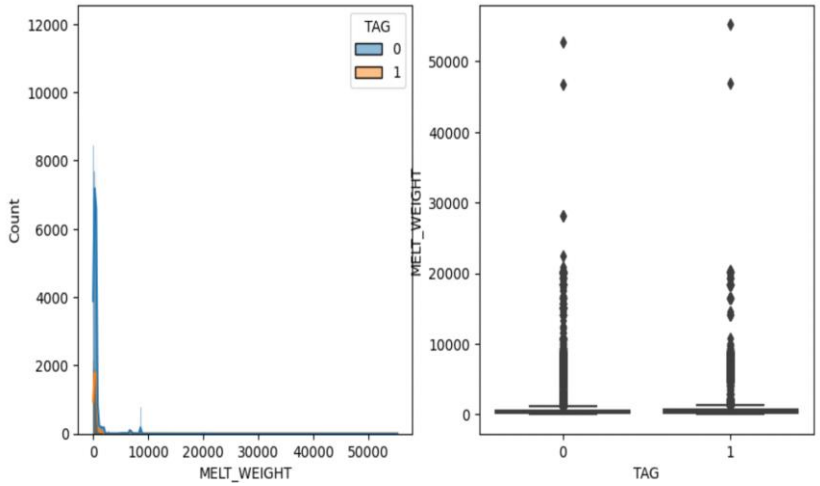
- MELT_TEMP



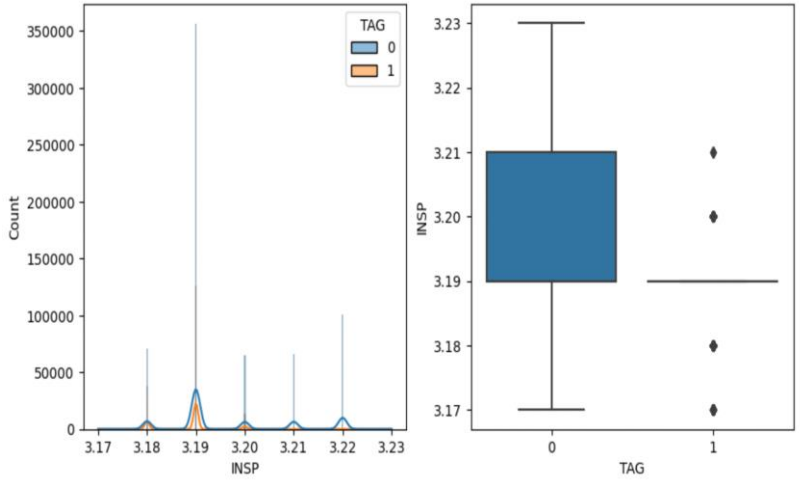
- MOTORSPEED



- MELT_WEIGHT



- INSP

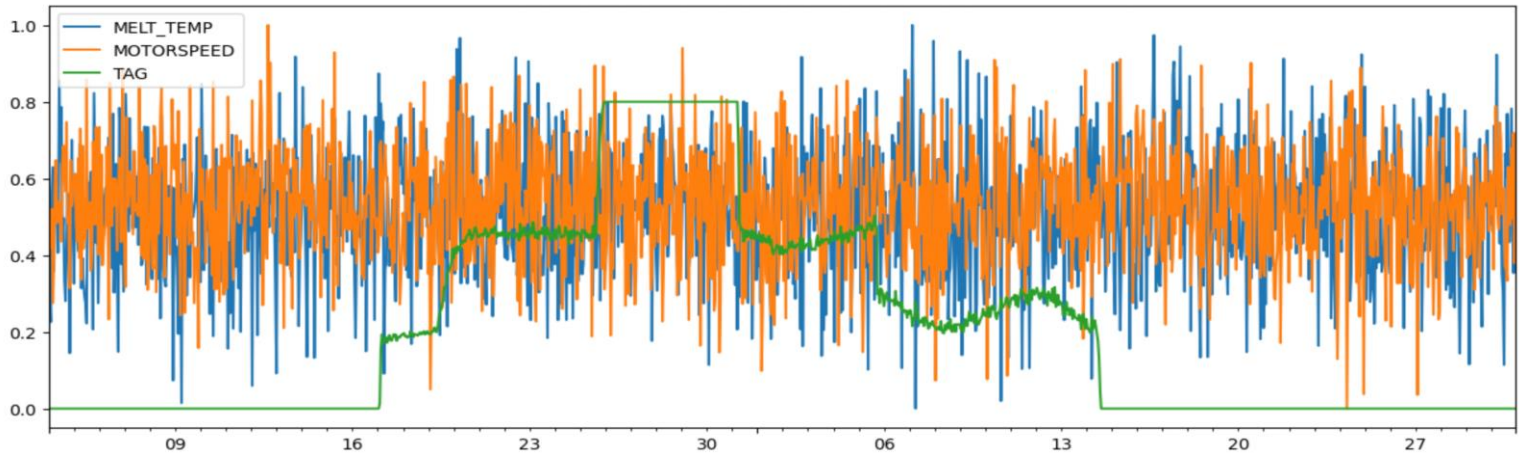


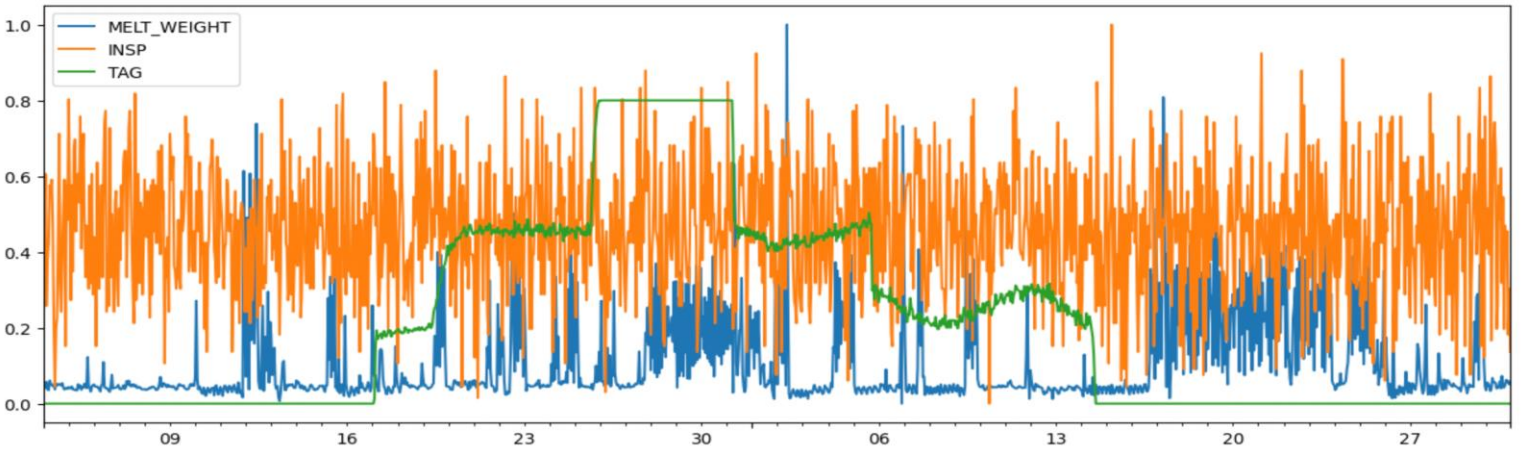
✓ MELT_TEMP와 MOTORSPEED에서 종속변수별 분포 차이가 나타남

2. 데이터 처리 과정

▪ 독립 및 종속변수 시계열 패턴 EDA

- DataFrame를 시간별로 리샘플링(평균처리)·정규화하여 시각화





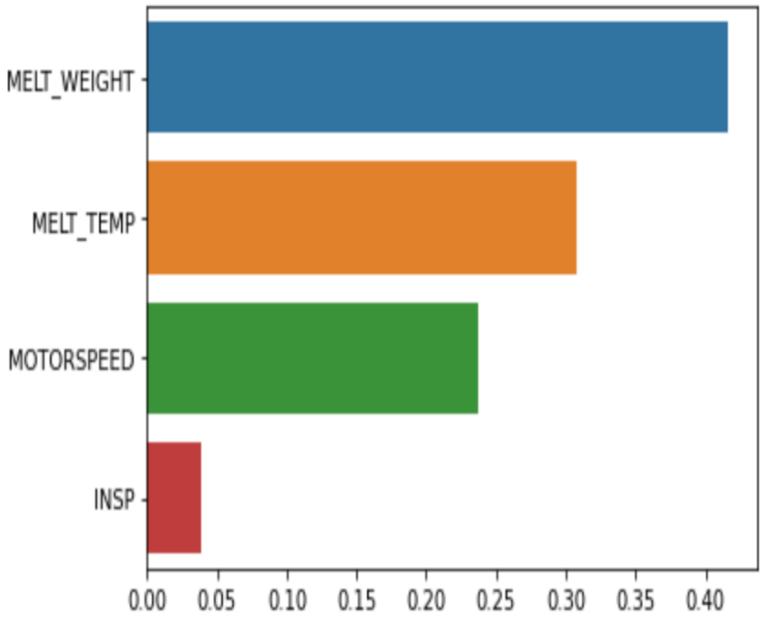
✓ 독립변수와 종속변수간의
뚜렷한 시계열 패턴 보이지 않음
(※ 시퀀스 모델 고려 제외)

3. 분석모델 개발

- (1차) Train set(70%) : Test set(30%)으로 구분 후 Train set으로 랜덤포레스트 모델링
- 불량 위험군 파생변수 생성 인사이트 도출

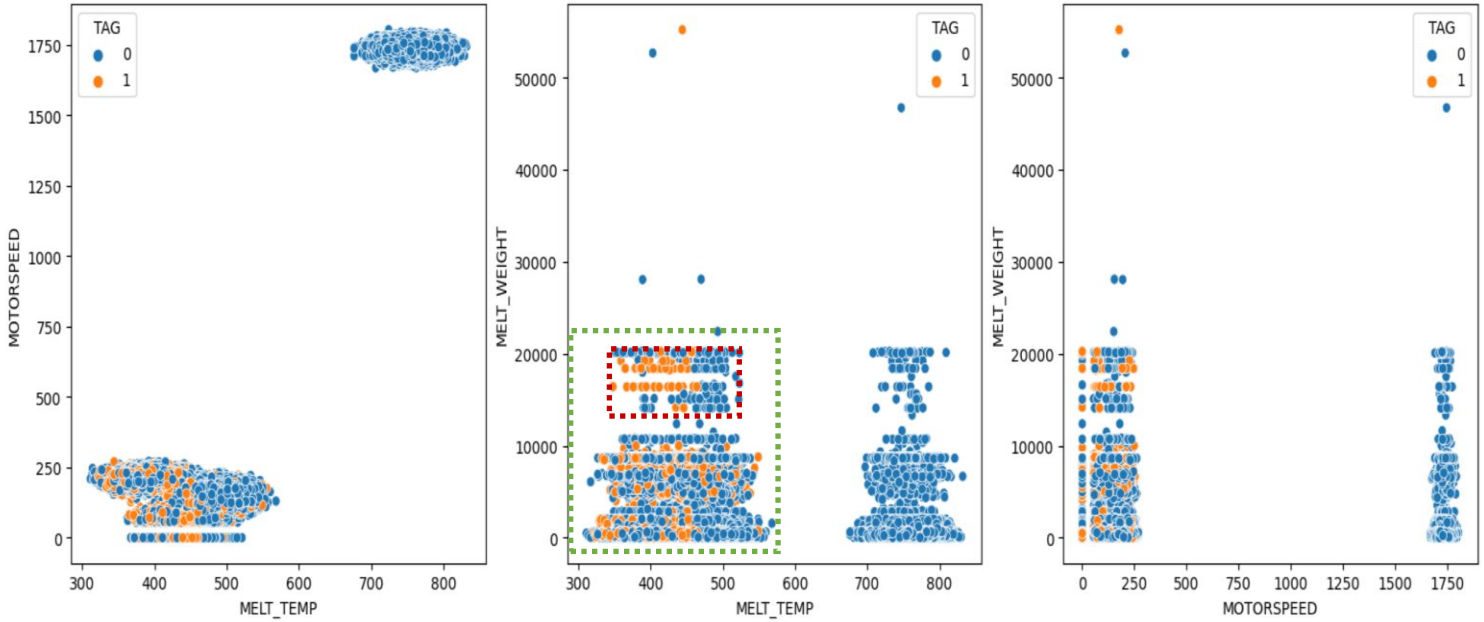
(랜덤포레스트) 피처 중요도 시각화

✓ MELT_WEIGHT : 높은 피처 중요도를 나타냄



- (각 축을 독립변수·색상을 종속변수로 설정) 산점도 시각화

✓ 불량 위험군 파생변수 생성 인사이트 도출
- 해당구역의 샘플은 정수 인코딩(경계군 : 0.5, 위험군 : 2)



※ (전체 데이터) 불량비율 0.21→ (경계군) 불량비율 0.27 → (위험군) 불량비율 0.50

3. 분석모델 개발

- (최종) Train set(70%) : Test set(30%)으로 구분 후 Train set으로 로지스틱회귀 모델링
- 불량 위험군 파생변수 생성 및 SMOTE(타겟변수 불균형 고려) 적용

(제공 데이터 - 정규화) 로지스틱회귀 모델링

✓ recall, F1 스코어가 매우 낮음

- Confusion Matrix

실제 \ 예측	0(OK)	1(NG)
0(OK)	193,922	3,518
1(NG)	50,837	2,283

- 평가지표

accuracy	0.7831
precision	0.3936
recall	0.0430
F1	0.0775

(SMOTE 적용) 로지스틱회귀 모델링

✓ recall, F1 스코어가 상당히 개선

- Confusion Matrix

실제 \ 예측	0(OK)	1(NG)
0(OK)	128,775	68,665
1(NG)	11,308	41,812

- 평가지표

accuracy	0.6808
precision	0.3785
recall	0.7871
F1	0.5112

(불량 위험군 변수 생성) 로지스틱회귀 최종모델링

✓ accuracy, recall이 소폭 개선

- Confusion Matrix

실제 \ 예측	0(OK)	1(NG)
0(OK)	128,783	68,657
1(NG)	11,305	41,815

- 평가지표

accuracy	0.6809
precision	0.3785
recall	0.7872
F1	0.5112

4. 분석결과 및 시사점

✓ 타겟변수 불균형 개선을 위한 SMOTE 및 불량 위험군 파생변수 생성의 효과성

- 최종 로지스틱 회귀 모델의 회귀계수 검증(statsmodel)

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.7430	0.009	191.227	0.000	1.725	1.761
MELT_TEMP	-1.5692	0.007	-223.786	0.000	-1.583	-1.555
MOTORSPEED	-0.5484	0.009	-64.472	0.000	-0.565	-0.532
MELT_WEIGHT	0.7310	0.021	34.105	0.000	0.689	0.773
INSP	0.1239	0.006	19.438	0.000	0.111	0.136
warning_area	-3.1222	0.032	-99.080	0.000	-3.184	-3.060

✓ 타 모델 대비 recall 및 F1 스코어에서 좋은 성능을 보이는 로지스틱 회귀

- (동일한 실험 조건) 랜덤포레스트와의 비교

(로지스틱회귀)

accuracy	0.6809
precision	0.3785
recall	**0.7872**
F1	**0.5112**

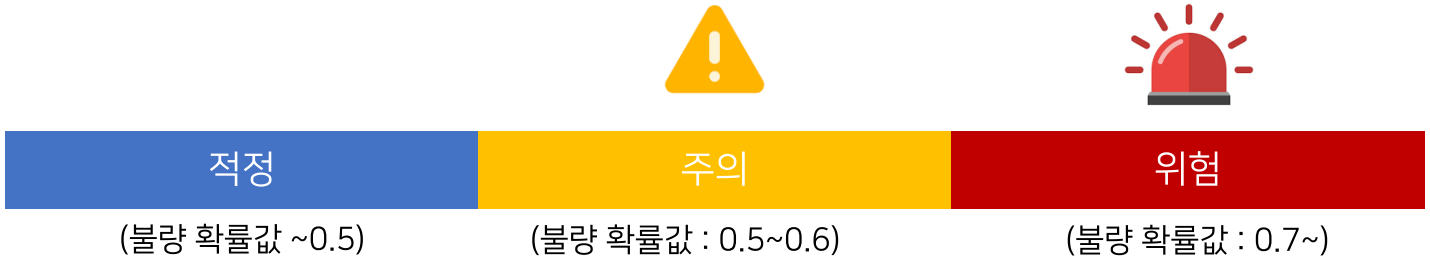
(랜덤포레스트)

accuracy	**0.7344**
precision	**0.3957**
recall	0.4799
F1	0.4338

5. 모델 제안 및 고도화 방안

- (모델 제안) SMOTE·불량 위험군 파생변수를 활용한 로지스틱회귀 불량 예측시스템

- 로지스틱회귀 불량 예측시스템 예시 – 작업장 패널 등에 안내 디스플레이 부착



✓ Why Logistic Regression?

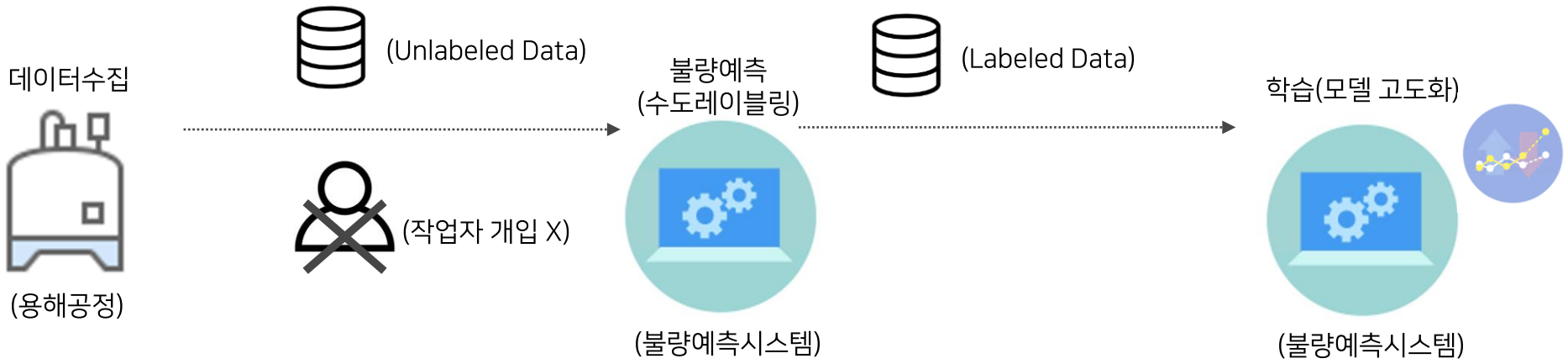
- **빠른 컴퓨팅 연산 속도 활용** : 6초에 한번씩 체크되는 공정 센서에 적용 가능
- **확률 모델 장점 사용** : 불량 확률 구간에 따른 위험도 판단 가능
- **모델 유연성 확보** : 임계치 조정으로 재현율 또는 정밀도에 초점을 맞춘 예측시스템 개발 가능

5. 모델 제안 및 고도화 방안

- (고도화 방안) 수도레이블링을 통한 모델 고도화 프로세스 구축

※ 추가적인 공정 데이터를 활용하는 프로세스를 통해 불량 예측시스템을 더 강건하게 만들고자 하는 목표

- 수도레이블링 적용 예시 – 공정에서 발생하는 데이터를 작업자 개입없이 라벨링



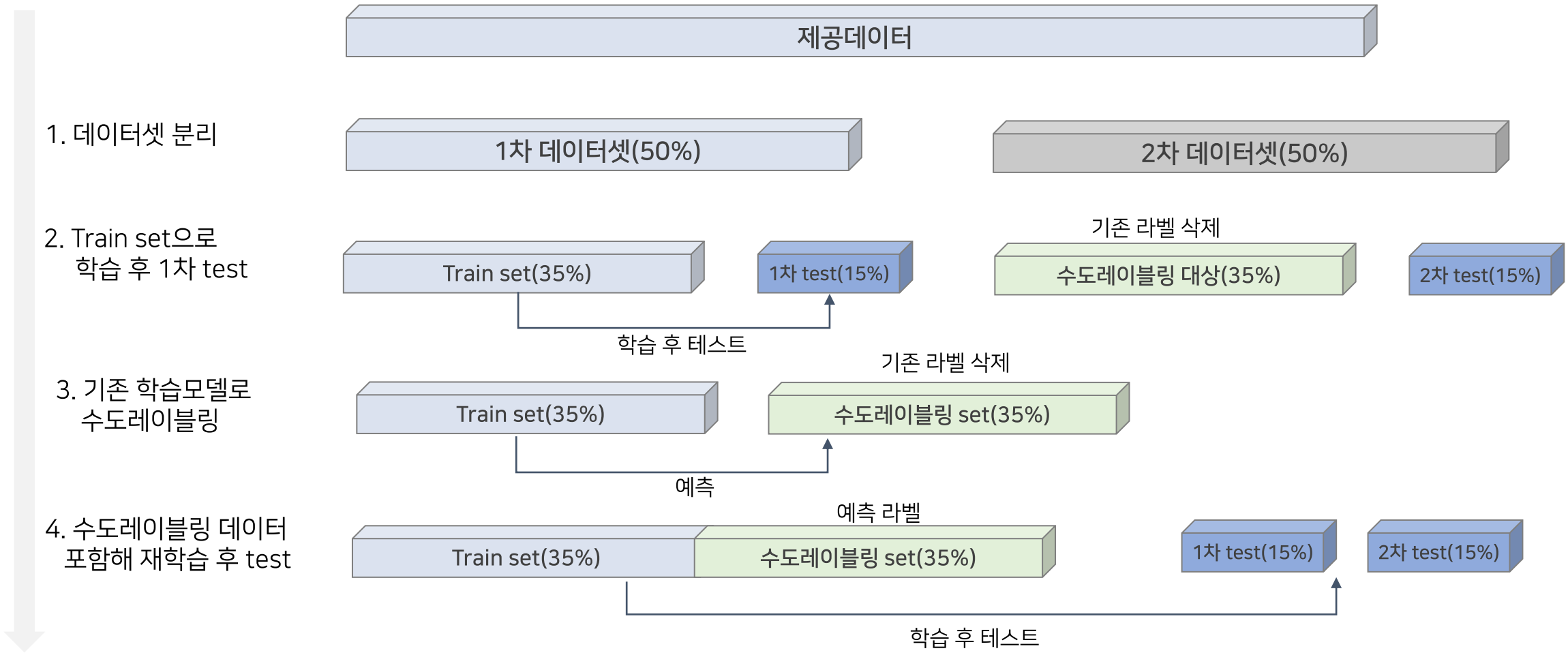
✓ Why pseudo-labeling?

- **수집 데이터 라벨링 비용 절감** : 작업자가 불량 여부를 판단하는 기존의 높은 라벨링 비용 개선
- **더 많은 공정 데이터 활용 가능** : 라벨링되지 않아 사용하지 못했던 데이터를 활용하여 공정 최적화 가능
- **로지스틱회귀 모델과의 시너지 효과** : 임계치 조정으로 불량 예측에 더 강건한 모델 개발 가능

5. 모델 제안 및 고도화 방안

- (고도화 방안) 수도레이블링을 통한 모델 고도화 프로세스 시뮬레이션

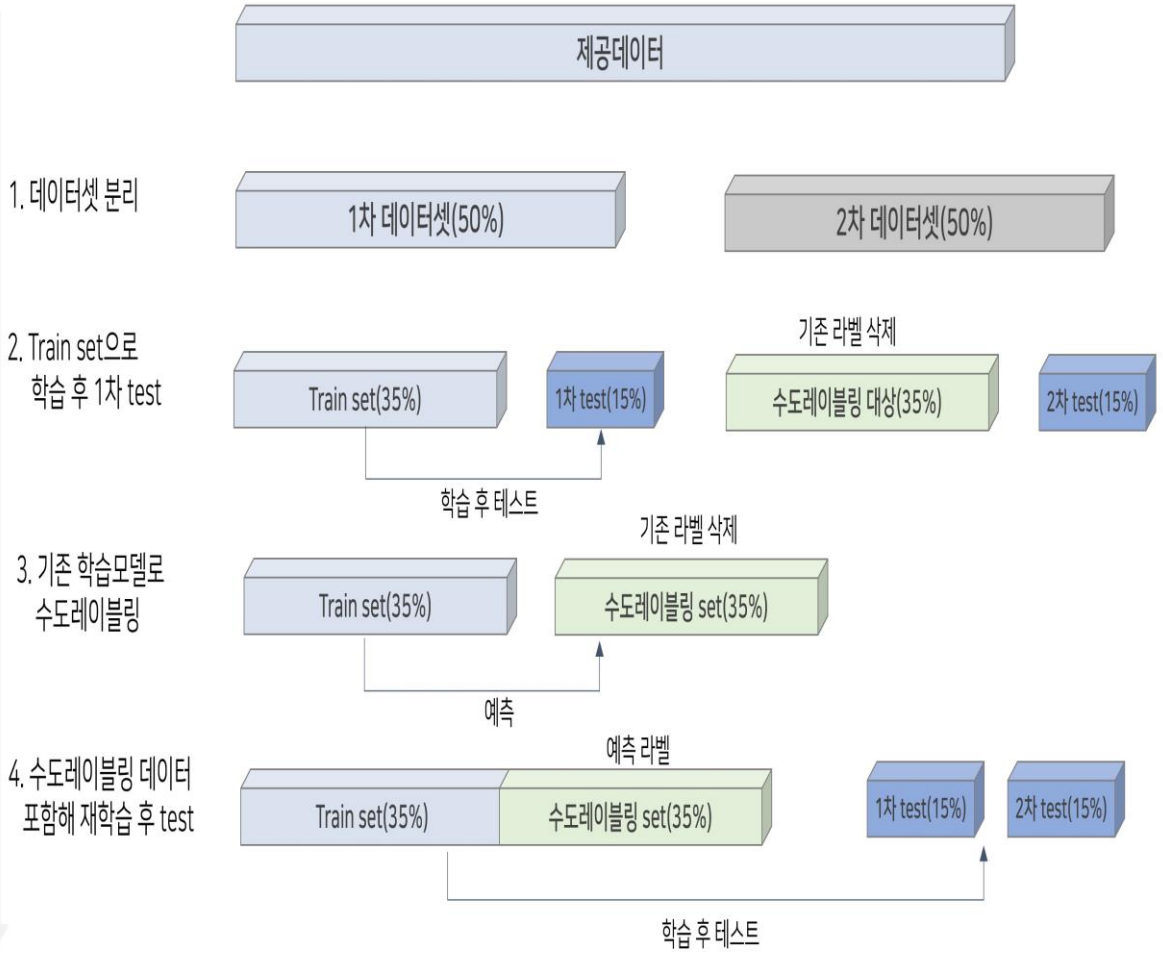
• 시뮬레이션 수행방법



5. 모델 제안 및 고도화 방안

■ (고도화 방안) 수도레이블링을 통한 모델 고도화 프로세스 시뮬레이션 평가

- 시뮬레이션 수행방법



- 시뮬레이션 수행 평가

✓ 수도레이블링을 통해 (precision 감소 대비) recall이 개선되는 효과

※ 임계치 조정으로 특정 지표에 초점을 맞춘 모델 개발 가능하며, 추가적인 수집 데이터를 활용한다면 성능 향상 기대

- 1차 Test set 비교

	(Train set)
accuracy	**0.6970**
precision	**0.4860**
recall	0.7780
F1	**0.5983**

	(Train + 수도레이블링 set)
accuracy	0.6677
precision	0.4593
recall	**0.8221**
F1	0.5893

- 2차 Test set 비교

	(Train set)
accuracy	**0.6409**
precision	**0.2674**
recall	0.9729
F1	**0.4195**

	(Train + 수도레이블링 set)
accuracy	0.5916
precision	0.2458
recall	**0.9970**
F1	0.3943

5. 모델 제안 및 고도화 방안

✓ SMOTE·불량 위험군 파생변수를 활용한 불량 예측시스템

“제조공정 자원 최적화를 위한 불량 예측시스템의 오늘과 내일”

✓ 수도레이블링을 통한 모델 고도화방안

6. 파급효과

✓ SMOTE·불량 위험군 파생변수를 활용한 불량 예측시스템

- 불량제품 사전 방지
- 불량 오인지에 따른 불필요한 공정 가동 감소

✓ 수도레이블링을 통한 모델 고도화

- 데이터 수집을 위한 인건비 절감
- 데이터 자원 활용 극대화

자원 최적화

감사합니다.