2023년 제 3회 K-인공지능 제조데이터 분석 경진대회

열처리 공정 품질 예측을 위한 가중 평균 앙상블 모델 개발

팀 BAA | 김소연 박시윤 최예진

Contents

개요

열처리 공정

- 금속 재료, 기계 부품, 금형 공구의 기계적 성질을 변화시키기 위해 가열과 냉각을 반복함으로써 유용한 성질 (내마모성, 내충격성, 수명연장) 등을 부여하는 공정
- 제조 공정의 중간 또는 최종단계에서 주로 운용됨
- 주로 경화와 인성을 향상시키기 위해 수행되며, **목표하는 물성을 얻는 것이 열처리 공정의 목표**임
- **금속 열처리는 같은 조건으로 공정을 적용해도 최종 제품의 물성이 달라지는 문제가** 있기 때문에 물성이 달라지는 **원인을 파악하면 제품의 수율을 안정시킬** 수 있는 특징을 가짐

- 본 과제의 열처리 공정은 전기 가열식 설비를 사용하며, **소입 → 염욕 → 세정 → 건조** 과정을 거침
- 소입로에서 가스 침탄과 오스테나이트화가 이루어진 후, 염욕 단계에서 250°C ~ 450°C의 냉각을 진행하여 베이나이트 조직을 얻는 것을 목표로 함

열처리 데이터셋

공정 데이터	
Row	2,939,722 개
Column	21개
Total	61,734,162개

- 제조 분야 : 자동차 부품 (에어백, 안전벨트)
- 수집 장비 : Manufacturing Execution System (MES)
- 수집 기간 : 2022년 01월 ~ 2022년 07월 (총 6개월)
- 수집 주기 : 배치별 사이클 타임 (약 1초)

- Raw Data 컬럼별 평균값

건조 1존 OP	건조 2존 OP	건조로 온도 1 Zone	건조로 온도 2 Zone	세정기	소입 1존 OP	소입 2존 OP	소입 3존 OP	소입 4존 OP	소입로 CP 값	소입로 CP 모니터 값	소입로 온도 1 Zone	소입로 온도 2 Zone	소입로 온도 3 Zone	소입로 온도 4 Zone	솔트 컨베이어 온도 1 Zone	솔트 컨베이어 온도 2 Zone	솔트조 온도 1 Zone	솔트조 온도 2 Zone
69.89	20.45	100.01	100.02	67.72	75.64	54.86	53.86	71.09	0.45	0.0	859.21	860.0	860.0	860.01	284.0	279.93	331.81	332.18

열처리 데이터셋

품질 데이터	
Row	136 개
Column	7 개
Total	952 개

- 배정번호, 작업일, 공정명, 설비명, 양품수량, 불량수량, 총수량
- 7개 컬럼 중 공정명, 설비명 컬럼은 공통 데이터의 반복임
- 불량수량 / 총수량 을 사용해서 공정별 불량률 계산이 가능함

- 품질 데이터

	배정번호	작업일	공정명	설비명	양품수량	불량수량	총수량
0	102410	2022-01-03	열처리	열처리 염욕_1	15160	3	15163
1	102585	2022-01-03	열처리	열처리 염욕_1	29892	10	29902
2	102930	2022-01-04	열처리	열처리 염욕_1	59616	30	59646
3	103142	2022-01-05	열처리	열처리 염욕_1	74730	13	74743
4	103675	2022-01-06	열처리	열처리 염욕_1	14979	2	14981
...	...	...	...	...	...	...	...

EDA

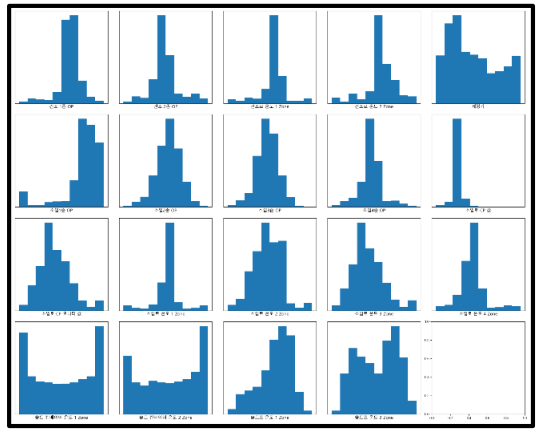
결측치 비율

배정번호별 결측치 비율 확인

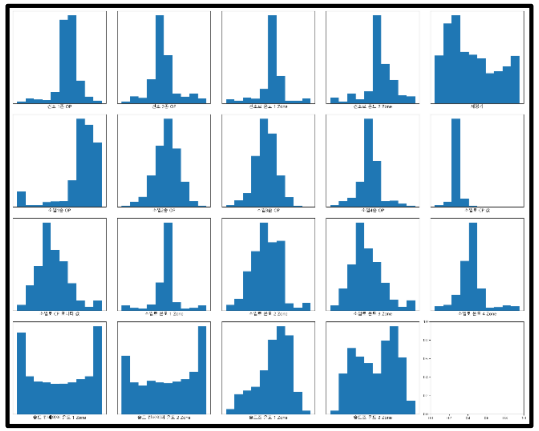
Max	0.087
Min	0.0

→ 최대 결측비율이 **0.087**로 적기 때문에 **결측치는 제거**하는 것으로 결정

배정번호별 컬럼별 히스토그램



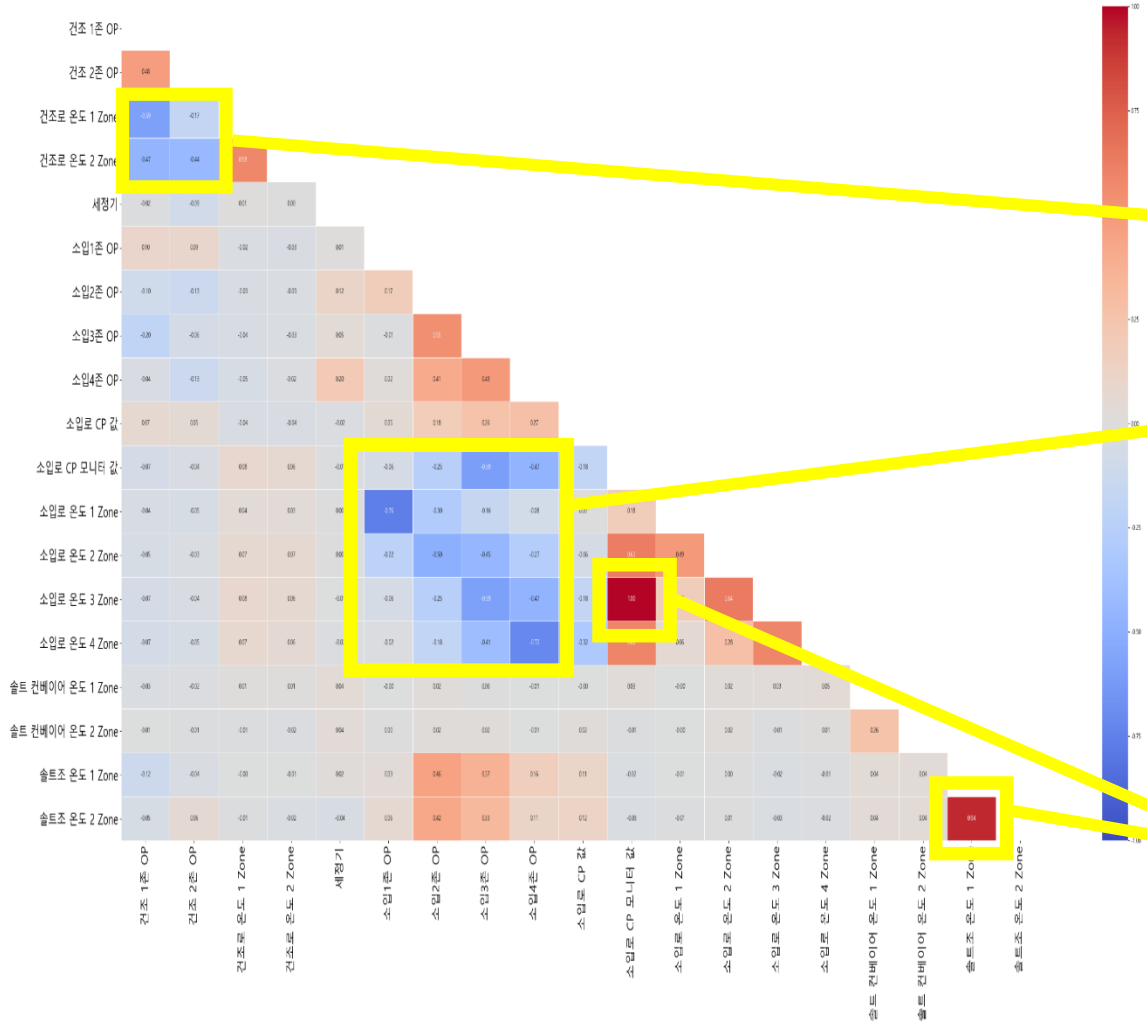
배정번호 : 102410



배정번호 : 102585

- 모든 컬럼이 정규 분포 모양은 아닌 것으로 확인됨

상관관계 분석 - Heat map



- OP 컬럼 - 온도 컬럼 : **비교적 큰 값의 음의 상관계수**를 가짐

OP	온도	Corr
건조 1존 OP	건조로 온도 1 Zone	- 0.6
건조 2존 OP	건조로 온도 2 Zone	- 0.4
소입 1존 OP	소입로 온도 1 Zone	- 0.8
소입 2존 OP	소입로 온도 2 Zone	- 0.5
소입 3존 OP	소입로 온도 3 Zone	- 0.6
소입 4존 OP	소입로 온도 4 Zone	- 0.7

- 그 외 **상관계수가 높게 나타난** 컬럼

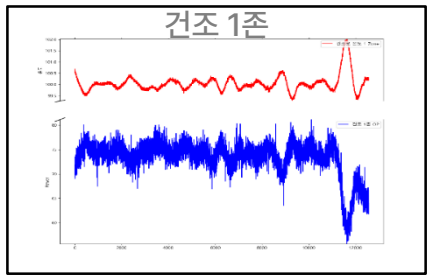
소입로 CP 모니터 값	소입로 온도 3 Zone	1.0
솔트조 온도 1 Zone	솔트조 온도 2 Zone	0.9

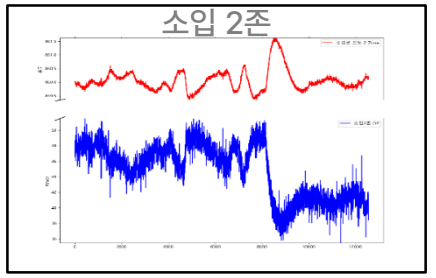
동일 공정의 OP컬럼 – 온도컬럼

> 배정번호별 공정 시간을 x축으로 두고
> 음의 상관계수가 크게 나타난 두 컬럼을 시각화
> → **음의 상관관계 확인**

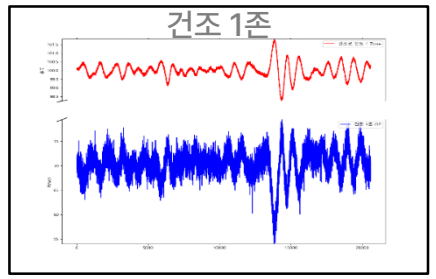
- 각 OP 값은 온도를 유지하기 위한 출력량으로,
  각 Zone의 온도와 관계있는 것으로 확인됨

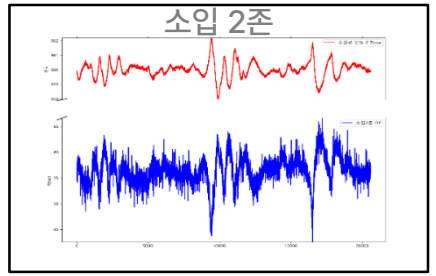
배정번호 :104126





배정번호 :135615





Red - 온도
Blue - OP

다중공선성 해결을 위한 변수 제거

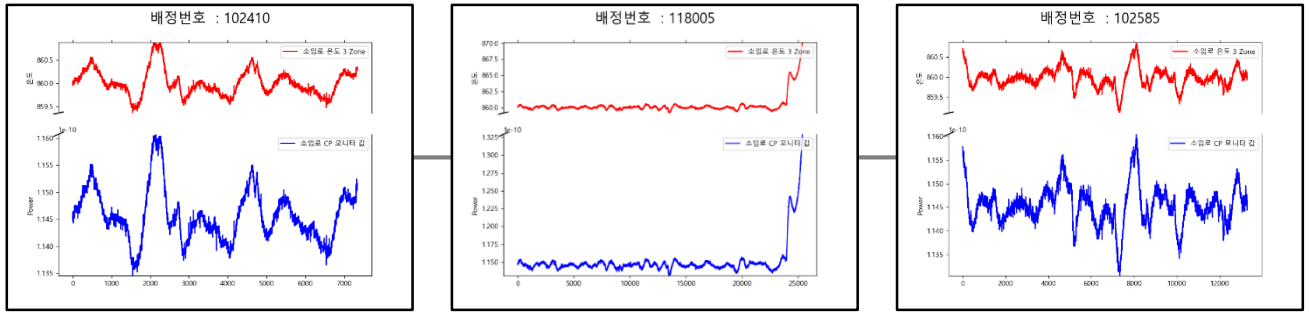
- 열처리 공정에서는 **출력량 보다 온도가 품질에 더 큰 영향**을 미치므로, **OP 변수를 제거하기로 결정**

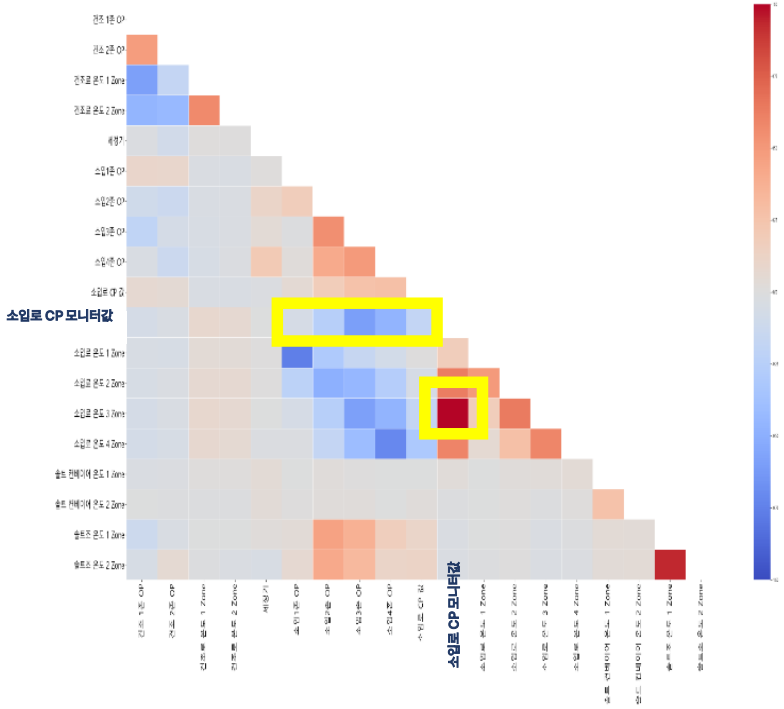
OP	온도
건조 1존 OP	건조로 온도 1 Zone
건조 2존 OP	건조로 온도 2 Zone
소입 1존 OP	소입로 온도 1 Zone
소입 2존 OP	소입로 온도 2 Zone
소입 3존 OP	소입로 온도 3 Zone
소입 4존 OP	소입로 온도 4 Zone

소입로 온도 3 Zone – 소입로 CP 모니터 값

상관계수 : 1.0 로 크게 나타난 두 컬럼 시각화 → **양의 상관관계 확인**

소입로 온도 3 Zone / 소입로 CP 모니터 값





다중공선성 해결을 위한 변수 제거

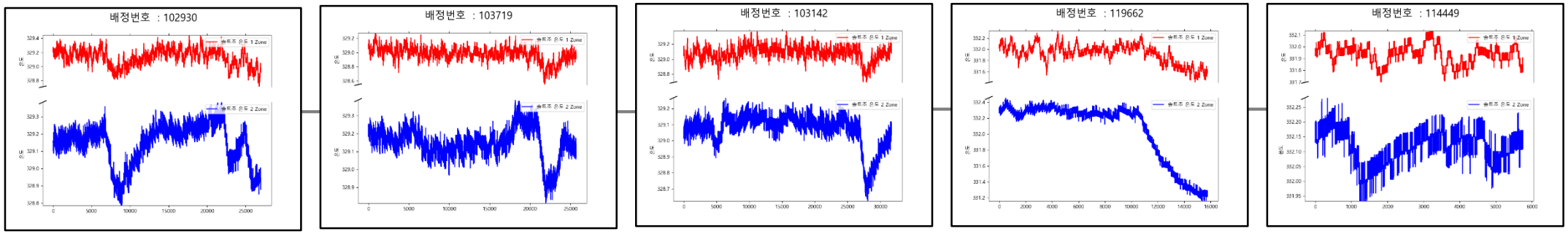
- 소입로 CP 모니터 컬럼은 소입로 온도 3 Zone 이외의 컬럼들과도 비교적 높은 상관관계를 가짐.
- 소입로 CP 모니터 값 컬럼과 소입로 3 Zone 온도 컬럼 두 개 중, 품질에 더 큰 영향을 미치는 온도 컬럼을 사용하는 것으로 결정함

컬럼 1	컬럼 2	Corr
~~소입로 CP 모니터 값~~	소입로 온도 3 Zone	1.0

솔트조 온도 1 Zone – 솔트조 온도 2 Zone

상관계수 : 0.9 → **양의 상관관계 확인**

솔트조 온도 1 Zone / 솔트조 온도 2 Zone



열처리 공정에서 중요한 온도 변수

- **열처리 공정에서 솔트조 온도는 오스테나이트를 냉각하여 베이나이트 형성을 하기 위한 중요 인자임**
- 솔트조 1,2 Zone 온도 컬럼의 시각화 그래프 및 변동계수 확인결과
- 상관관계는 높으나, 솔트조 2 Zone의 온도 변동폭 및 변동 계수가 가 큰 것을 확인하여 변수를 제거하지 아니함

컬럼명	변동계수(CV)
솔트조 온도 1 Zone	0.2
솔트조 온도 2 Zone	0.3

데이터 전처리

데이터 전처리

※ 앞에서 언급한 OP 컬럼들(6개) 과 소입로 CP 모니터 컬럼(1개) 은 삭제함

전체 컬럼 개수 : 14개

data.csv

결측치 삭제 | 2939722 → 2939241

```
Index: 2939241 entries, 2 to 2939721
Data columns (total 14 columns):
 #   Column              Dtype
---  ------              -----
 0   TAG_MIN             datetime64[ns]
 1   배정번호                int64
 2   건조로 온도 1 Zone      float64
 3   건조로 온도 2 Zone      float64
 4   세정기                 float64
 5   소입로 CP 값           float64
 6   소입로 온도 1 Zone      float64
 7   소입로 온도 2 Zone      float64
 8   소입로 온도 3 Zone      float64
 9   소입로 온도 4 Zone      float64
 10  솔트 컨베이어 온도 1 Zone  float64
 11  솔트 컨베이어 온도 2 Zone  float64
 12  솔트조 온도 1 Zone      float64
 13  솔트조 온도 2 Zone      float64
dtypes: datetime64[ns](1), float64(12), int64(1)
```

TAG_MIN 컬럼 타입 변환 | object → datetime

```
---  ------              -----
 0   TAG_MIN             datetime64[ns]
 1   배정번호                int64
 2   건조로 온도 1 Zone      float64
 3   건조로 온도 2 Zone      float64
 4   세정기                 float64
 5   소입로 CP 값           float64
 6   소입로 온도 1 Zone      float64
 7   소입로 온도 2 Zone      float64
 8   소입로 온도 3 Zone      float64
 9   소입로 온도 4 Zone      float64
 10  솔트 컨베이어 온도 1 Zone  float64
 11  솔트 컨베이어 온도 2 Zone  float64
 12  솔트조 온도 1 Zone      float64
 13  솔트조 온도 2 Zone      float64
dtypes: datetime64[ns](1), float64(12), int64(1)
```

data.csv

배정번호별 그룹화 | mean, std, min, max

- **열처리 공정 특성상, 재료의 성질을 일정하게 만드는 것이 중요하다고 판단하여 통계값을 사용함**

통계값	사용 이유
Mean	대량의 데이터를 요약하는 대표값
Std	공정 중 발생하는 변동이 품질에 미치는 영향평가
Min & Max	공정 중 발생할 수 있는 극한 조건에서의 공정 파악

- 배정번호를 기준으로 그룹화를 진행하여 통계 변수를 생성함
- 각 변수 컬럼당 Mean, Std, Min, Max 총 4개의 변수 생성
- 총 컬럼의 개수 변화 : 14개 → 50개

배정번호	건조로 온도 1 Zone_min	건조로 온도 1 Zone_max	건조로 온도 1 Zone_mean	건조로 온도 1 Zone_std	...	솔트조 온도 2 Zone_mean	솔트조 온도 2 Zone_std	솔트조 온도 2 Zone_min	솔트조 온도 2 Zone_max
102410	97.83	101.755	99.943	0.593	...	329.07	0.1165	328.8	329.300
102585	97.98	101.760	99.987	0.515	...	328.92	0.0891	328.6	329.121
102930	97.77	101.999	99.995	0.472	...	329.14	0.1155	328.7	329.383
103142	98.31	101.818	100.00	0.331	...	329.07	0.1004	328.6	329.267
103675	97.72	102.076	99.983	0.655	...	329.11	0.0788	328.9	329.320
...	...	...	...	...	...	...	...	...	...

정규화 | Minmax Scaler

후술할 인공신경망 모델링을 위해
데이터 Minmax Scaler 진행

배정번호	건조로 온도 1 Zone_min	건조로 온도 1 Zone_max	건조로 온도 1 Zone_mean	건조로 온도 1 Zone_ std	...	솔트조 온도 2 Zone_mean	솔트조 온도 2 Zone_std	솔트조 온도 2 Zone_ min	솔트조 온도 2 Zone_max
102410	97.83	101.755	99.943	0.593	...	329.07	0.1165	328.8	329.300
102585	97.98	101.760	99.987	0.515	...	328.92	0.0891	328.6	329.121
102930	97.77	101.999	99.995	0.472	...	329.14	0.1155	328.7	329.383
103142	98.31	101.818	100.00	0.331	...	329.07	0.1004	328.6	329.267
103675	97.72	102.076	99.983	0.655	...	329.11	0.0788	328.9	329.320
...	...	...	...	...	...	...	...	...	...

인공신경망 모델의 경우, 신경망의 가중치 초기화 및 활성화 함수가 특정 범위의 입력에 최적화되어 있기 때문에, 각 컬럼별로 Min Miax 정규화를 적용하여 활성화 함수의 포화와 경사하강법의 불안정성 문제를 방지하고자함

102410	0.72682	0.75807	0.23664	0.7479	...	0.0373	0.1876	0.162739	0.04411
102585	0.78259	0.64244	0.30600	0.7497	...	0.0000	0.1235	0.136950	0.00000
102930	0.79278	0.57955	0.20603	0.8340	...	0.0573	0.1854	0.158737	0.06456
103142	0.80495	0.37151	0.46269	0.7701	...	0.0380	0.1501	0.122277	0.03597
103675	0.77772	0.84856	0.18420	0.8612	...	0.0484	0.0994	0.183859	0.04903
...	...	...	...	...	...	...	...	...	...

quality.xlsx

반복 및 관련없는 변수 삭제 | 공정명, 설비명, 작업일

	배정번호	작업일	공정명	설비명	양품수량	불량수량	총수량
0	102410	2022-01-03	열처리	열처리 염욕_1	15160	3	15163
1	102585	2022-01-03	열처리	열처리 염욕_1	29892	10	29902
2	102930	2022-01-04	열처리	열처리 염욕_1	59616	30	59646
3	103142	2022-01-05	열처리	열처리 염욕_1	74730	13	74743
4	103675	2022-01-06	열처리	열처리 염욕_1	14979	2	14981
...	...	...	...	...	...	...	...

불량률 파생변수 추가 | 불량수량 / 총수량

	배정번호	양품수량	불량수량	총수량	불량률
0	102410	15160	3	15163	0.020
1	102585	29892	10	29902	0.033
2	102930	59616	30	59646	0.050
3	103142	74730	13	74743	0.017
4	103675	14979	2	14981	0.013
...	...	...	...	...	...

배정번호 기준으로 데이터 합치기

[data.csv + quality.xlsx]

배정번호	건조로 온도 1 Zone_ min	건조로 온도 1 Z one_m ax	…	솔트조 온도 2 Zone _min	솔트조 온도 2 Zone _max
102410	0.726	0.758	…	0.162 739	0.04 411
102585	0.782	0.642	…	0.136 950	0.00 000
102930	0.792	0.579	…	0.158 737	0.06 456
103142	0.804	0.371	…	0.122 277	0.03 597
103675	0.777	0.848	…	0.183 859	0.04 903
…	…	…	…	…	…

+

배정번호	…	총수량	불량률
102410	…	0.162 739	0.044 11
102585	…	0.136 950	0.000 00
102930	…	0.158 737	0.064 56
103142	…	0.122 277	0.035 97
103675	…	0.183 859	0.049 03
…	…	…	…

→

배정번호	건조로 온도 1 Zone_ min	건조로 온도 1 Z one_m ax	…	총수량	불량률
102410	0.726	0.758	…	0.162 739	0.0441
102585	0.782	0.642	…	0.136 950	0.0000
102930	0.792	0.579	…	0.158 737	0.0645
103142	0.804	0.371	…	0.122 277	0.0359
103675	0.777	0.848	…	0.183 859	0.0490
…	…	…	…	…	…

데이터 전처리

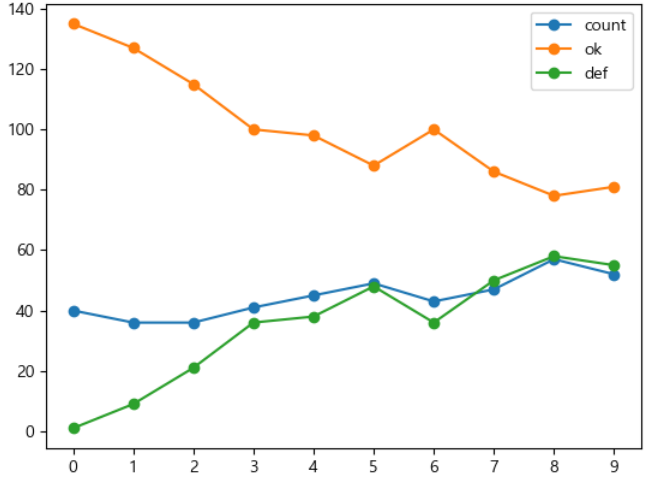
Auto Labeling

머신러닝 알고리즘 및 통계적 기법을 사용하여 레이블을 할당하거나 생성하는 프로세스

Unlabled Data	5% Labeling	학습	95% Labeling
라벨링이 필요한 데이터	Unlabled Data 의 5% 에 라벨링 직접 부여	데이터 특성 학습 모델 생성	Unlabled Data 의 95% 에 대한 Labeling 값 추론

불량률 데이터	위험 / 안정 라벨링	오토 라벨링 모델 생성	Auto Labeling
배정 번호 136개에 해당하는 공정의 불량률 컬럼	불량률을 기준으로 상위 4개를 위험(0)으로, 하위 7개를 안정(1)으로 지정	데이터의 특성과 패턴을 학습하여 분류, 군집화, 회귀 및 이상치 탐지 분석	데이터를 자동으로 레이블하여 진행 공정별 특성에 알맞은 라벨링 진행 가능

안정 / 위험 개수 지정 기준



1. 첫째, 불량률이 높은 순을 기준으로 하위 70%를 안정(1)으로 라벨링한 경우와 차이가 작을 것
2. 둘째, 안정과 위험의 개수 비율이 안정적일 것

→ 기존과의 차이는 41개, 위험으로 분류한 경우는 36개, 안정으로 분류한 경우는 100개

모델 학습에 들어가는 Count(41개), 위험(36개), 안정(100개)의 데이터 수가 적절하고 Model의 성능이 가장 잘 나오는 70%_7_4 데이터 셋 선정

Modeling

머신 러닝과 딥러닝을 함께 사용하는 이유

- 딥러닝 모델은 특정 패턴이나 노이즈에 과적합 될 수 있는 경우가 존재하지만
  머신러닝 모델은 더 단순한 패턴을 학습하므로, 특정 노이즈를 무시하고 주요 패턴에 집중할 수 있음
- 따라서 두 모델이 서로 다른 오류 패턴을 가질 때, 둘을 조합하면 더 안정적인 예측이 가능할 것으로 판단함

독립변수 / 종속변수 정의

- 독립변수 : 총 48개의 컬럼
- 종속변수 : Auto Labeling 을 통해 생성한 공정의 '위험'/ '안정' 라벨링을 각각 0,1로 매핑한 데이터

배정번호	건조로 온도 1 Zone_min	건조로 온도 1 Zone_max	건조로 온도 1 Zone_mean	건조로 온도 1 Zone_std	…	솔트조 온도 2 Zone_mean	솔트조 온도 2 Zone_std	솔트조 온도 2 Zone_min	솔트조 온도 2 Zone_max	공정상태
102410	97.83	101.755	99.943	0.593	…	329.07	0.1165	328.8	329.300	1
102585	97.98	101.760	99.987	0.515	…	328.92	0.0891	328.6	329.121	1
…	…	…	…	…	…	…	…	…	…	…

RFE – 주요 컬럼 추출

- xgb = XGBClassifier(eval_metric='error', learning_rate= 0.06, n_estimators= 20)
- RFE(estimator= xgb, n_features_to_select=10, step=5, verbose=0)
- 총 48개 변수에서 중요도가 높은 10개 변수 추출

추출된 주요 변수 (10개)

- 건조로 온도 1 Zone_min
- 건조로 온도 1 Zone_std
- 건조로 온도 2 Zone_std
- 세정기_mean
- 소입로 온도 3 Zone_min
- 소입로 온도 3 Zone_std
- 소입로 온도 4 Zone_max
- 소입로 온도 4 Zone_std
- 솔트 컨베이어 온도 1 Zone_max
- 솔트조 온도 1 Zone_min

10개

배정번호	건조로 온도 1 Zone_min	건조로 온도 1 Zone_std	...	...	...	솔트 컨베이어 온도 1 Zone_max	솔트조 온도 1 Zone_min
102410	0.726	0.61984	...	...	...	0.0535	0.04411
102585	0.782	0.43919	...	...	...	0.0484	0.00000
102930	0.792	0.42066	...	...	...	0.0875	0.06456
103142	0.804	0.29120	...	...	...	0.0726	0.03597
103675	0.777	0.60181	...	...	...	0.0518	0.04903
...	...	...	...	...	...	...	...

XGBoost Modeling

Random Search 알고리즘으로 최적의 하이퍼파라미터 탐색

- 탐색할 하이퍼파라미터 범위 설정
- Param_dist = {'learning_rate' : [0.001, 0.01, ...], 'n_estimators' : [100, 200, ...], ...}

랜덤 서치 객체 생성

- RandomizedSearchCV

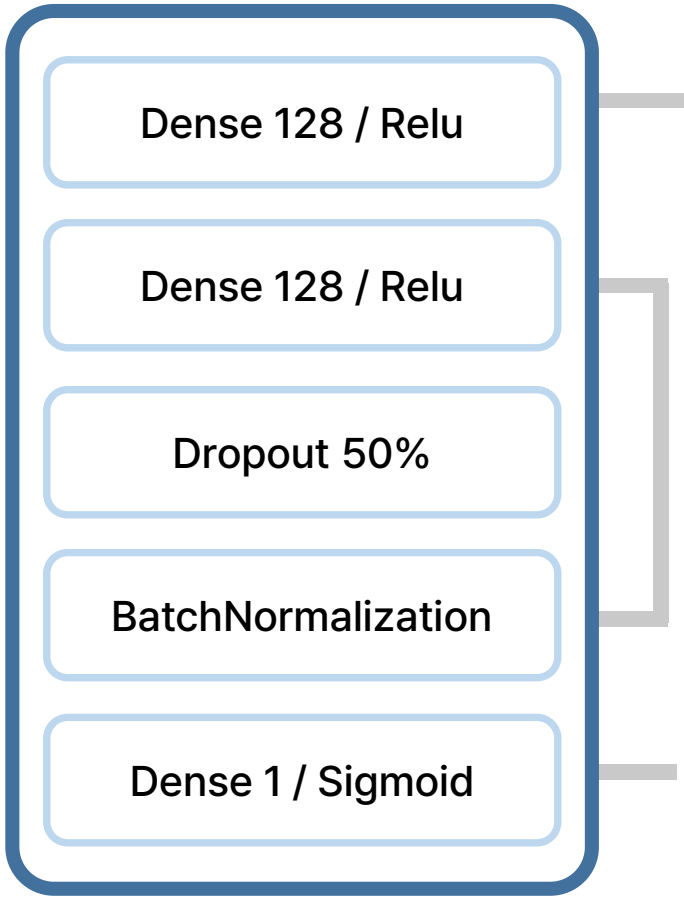
  (xgb, param_distributions = param_dist, n_iter = 10, cv = 5, scoring = 'accuracy', n_jobs = -1, random_state = 42)

Best Params	
Subsample	0.9
n_estimators	300
min_child_weight	5
max_depth	8
learning_rate	0.01
gamma	0.3
colsample_bytree	0.9

정확도 결과

훈련 정확도	0.934
검증 정확도	0.885
정확도 차	0.049

인공 신경망 (ANN)



Input Layer

활성화 함수 Relu를 이용

비선형성 추가

Hidden Layer

125개의 뉴런을 추가하여 복잡도 추가

과대적합 방지를 위하여 dropout 50% 추가

특정 뉴런에 의존하는 것을 방지하여 Dense 층 이후 추가

정규화 층을 dropout 이후에 추가하여 학습 안정성을 높임

내부 공선성 감소시키는 효과

Output Layer

활성화 함수 Sigmoid를 이용

이진분류

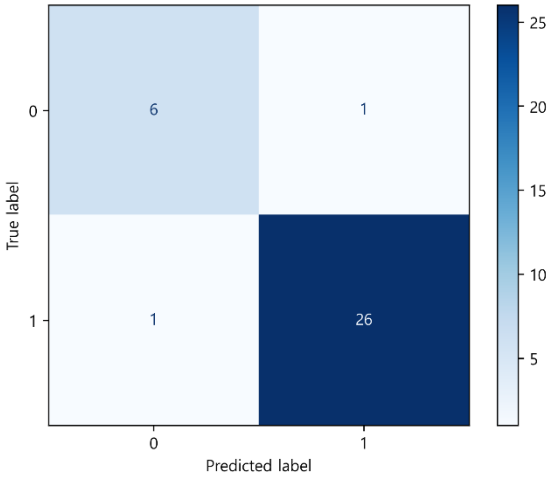
정확도 결과

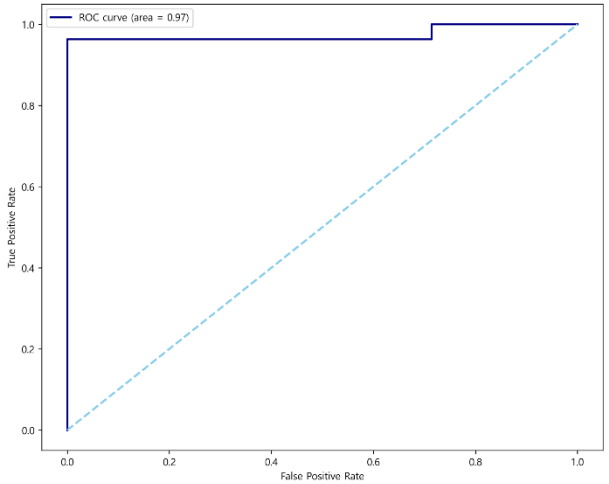
훈련 정확도	0.974
검증 정확도	0.923
정확도 차	0.011

Result

모델링 최종 결과

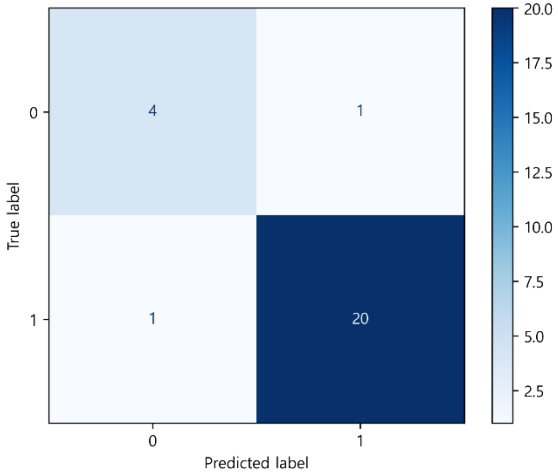
Test Data

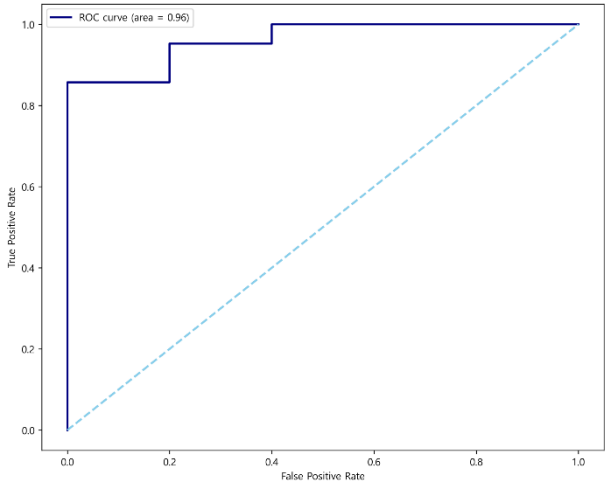




평가 지표	Test
Accuracy	0.9118
Recall	0.9630
F1 Scrore	0.9455
AUC value	0.9683

Validation Data





평가 지표	Validation
Accuracy	0.9231
Recall	0.9524
F1 Scrore	0.9524
AUC value	0.9524

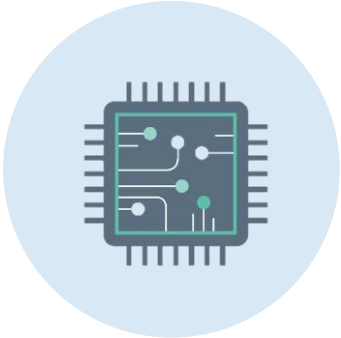
모델링 결과 시사점

- 본 모델은 전체 테스트 데이터에 대하여 약 92%의 높은 정확도로 분류함

- '위험'과 '안정'의 클래스 불균형을 고려하여 Recall, F1 Score, AUC 값을 추가 측정한 결과, 재현률(Recall)은 약 95.24%로 '위험' 공정을 높은 정확도로 인식하였으며, F1 Score와 AUC 값은 모두 0.9524로 위험 및 안정 공정을 구분하는 능력이 뛰어난 것으로 판단됨

- 이러한 결과는 모델 구조, 하이퍼파라미터 최적화, 주요 특성 도출 및 엔지니어링 과정을 통해 클래스 불균형 문제를 해결하고, 공정 효율을 향상시킬 수 있는 우수한 성능을 보였다는 것을 나타냄

기대효과



반도체 에칭 공정

반도체 에칭 공정은 열을 이용하여
표면 이물질 제거 및 패턴 형성
미세한 결함을 방지
품질 및 생산성 향상 가능

금속 가공 및 주조 공정

고온이나 고압 등 다양한 환경 조건에서의
불량을 사전에 예측하여 품질 수준 향상 기여

타 분야로의 확장

자동차 산업 뿐만아니라
철도 차량, 금형, 전기, 건설 중장비 등
광범위하게 활용 가능

기대효과

제품 품질 및 생산 효율 향상

제품 생산에 최적화된 공정 관리를 통해 제품의 품질을 높이고 불량률을 저하시켜 생산 효율 향상

불량품 처리 및 재작업 비용 감소

불량품 감소로 인한 불량품 처리 비용과 재작업에 소요되는 인력 및 원자재 비용을 절감할 수 있음

연속적 불량 생산 방지

실시간 모니터링 및 이상치 탐지를 통한 빠른 점검으로 연속적인 공정 불안 및 불량품 다량 생산을 방지함