제3회 K-인공지능 제조데이터 분석 경진대회 보고서

프로젝트명	**XAI를 이용한 열처리 공정 품질 이상 탐지 모델 해석**
팀명	**치킨마요덮밥**
내용요약	문제 정의 - 공정(설비) 개요 : 열처리 공정 - 이슈사항 : 열처리 공정은 연속공정으로 생산품이 불량으로 발견되어도 설비를 중단할 수 없으며, 생산품 모두 불량으로 배출될 경우가 발생한다. 제조데이터 및 처리과정 - 제조데이터 정의 : (공정 데이터), 행 2939722개, 열 21개, 총 데이터 61,734,162개 (품질 데이터), 행 137개, 열 7개, 총 데이터 959개 - 데이터 전처리 : 데이터 타입 변경, 결측치 대치, 변수 추가 및 제거, SMOTE-TOMEK 분석모델 개발 - AI 분석 방법론(알고리즘) : CatBoost, Shap Value 분석 결과 및 시사점 - 분석 결과 - 시사점 및 보완점 중소제조기업에 미치는 파급효과 - 파급효과

상기 본인(팀)은 위의 내용과 같이 제3회 K-인공지능 제조데이터 분석 경진대회 결과 보고서를 제출합니다.

2023 년 11월 3일

팀장 : 김창현 (서명)
팀원 : 김동준 (서명)
팀원 : 복현우 (서명)

한국과학기술원장 귀중

□ 문제정의

○ 공정(설비) 개요

– 열처리 공정

- 열처리 공정은 전기 가열식 설비로서 AUSTEMPER를 목적으로 설계되었으며 입구 PUSHER에서 로체로 장입한 후에 AUSTEMPER → 추출 → SALT 소입→ 추출 등의 작업이 자동으로 이루어지는 공정이다.
- 적당한 온도로 가열하여 안정된 Austenite 구역으로 유지한 후 Martensite 생성온도 이상의 염욕(약 250℃~ 450℃)속에 급랭시켜 Bainite 조직을 얻는 열처리 방법이다.
- Q/T처리 대신 신율, 단면 수축률, 충격치 등이 향상되고 인성이 우수해지며 Crack 및 변형이 감소된다.
- 또한, oil면 상의 압력을 조정하면서 Quenching을 행함으로 하나의 Quenching oil로 저온 oil부터 고온 oil까지 상당히 폭 넓은 범위의 냉각성을 발현한다.
- 흐름도 : 공급기 → 소입로 → 소입탱크 → Quenching(후세척) → 소려로 → 방청

○ 이슈사항

– 공정(설비)상의 문제 현황

- 열처리 공정은 생산품 투입부터 배출까지 약 2시간 이상 소요가 되며, 불량을 육안으로 확인하기 전까지 제품의 품질을 확인할 수 없다.
- 열처리 공정은 연속공정으로 생산품이 불량으로 발견되어도 설비를 중단할 수 없으며, 생산품 모두 불량으로 배출될 경우가 발생한다.

– 문제해결 장애요인

- 현장작업자가 불량을 검출하고 유형별로 수량을 입력하고 있지만, 불량 발생 시 정확한 원인을 파악하지 못하며 개선활동에도 어려움이 있다.
- 현장에서 발생하는 생산 데이터를 수집하고 실시간 모니터링을 하고 있지만 현재 이슈에 대한 파악만 가능하며, 원인분석 및 품질예측이 불가능하다.
- 현재 다양한 데이터가 수집되고 있으나 활용을 못하는 상황으로, 이는 대부분의 중소기업 제조 현장에서 공통으로 겪고 있는 것으로 판단된다.
- 새로운 공정이 시작될 때마다 온도 센서의 미작동/오작동 등으로 공정 초기 데이터가 제대로 수집되지 않고 있다.

– 극복 방안

- 공정 진행 중 생산품의 상태를 직접 확인/검사하지 못하는 현장 상황 등의 조건 속에서 공정 품질을 확보하기 위해서는 데이터에 기반을 둔 품질예측방법이 필요하다.
- 다양한 요인으로 인해 공정 중 실시간 변화하는 공정 데이터와 주요 품질검사항목의 결과값을 모델링하여 이를 바탕으로 생산 중에도 생산품의 품질을 예측하여 즉시 공정 제어에 이용할 수 있을 것이다.
- XAI 기법을 이용해 인공지능 모델의 결과를 해석할 수 있다. 즉, 생산품의 품질을 결정하는 데 영향을 미치는 주요 요인들과 그 중요도를 파악할 수 있다.

□ 제조데이터 정의 및 처리과정

○ 제조데이터 정의

– 데이터 유형/구조

- 공정 데이터(data.csv) : 행 2939722개, 열 21개, 총 데이터 61,734,162개
- 품질 데이터(열처리_품질데이터.xlsx) : 행 137개, 열 7개, 총 데이터 959개

– 데이터 변수/타입

- 공정 데이터

변수	설명	타입
TAG_MIN	데이터 습득 시간	object
배정번호	공정 작업 번호	int
건조 1~2존 OP	각 건조 온도 유지를 위한 출력 량(%)	float
건조로 온도 1~2 Zone	각 건조로 Zone 온도 값	float
세정기	세정기 온도 값	float
소입1~4존 OP	각 소입존 온도 유지를 위한 출력 량(%)	float
소입로 CP 값	침탄 가스의 침탄 능력의 량(%)	float
소입로 CP 모니터 값	소입로 출력 량에 대한 모니터 값	float
소입로 온도 1~4 Zone	각 소입로 Zone의 온도 값	float
솔트 컨베이어 온도 1~2 Zone	각 솔트 컨베이어 Zone의 온도 값	float
솔트조 온도 1~2 Zone	각 솔트조 Zone의 온도 값	float

<표 1> 공정 데이터 항목별 구성

- 품질 데이터

변수	설명	타입
배정번호	공정 작업 번호	int
작업일	공정 작업 일자	datetime
공정명	수행한 공정의 이름	object
설비명	공정을 진행한 설비의 이름	object
양품수량	공정에서 양품으로 분류된 생산품의 수량	int
불량수량	공정에서 불량으로 분류된 생산품의 수량	int
총수량	공정에서 생산된 총 수량	int

<표 2> 품질 데이터 항목별 구성

- 데이터 전처리

1. 데이터 타입 변경

- 공정 데이터의 TAG_MIN 변수는 데이터의 습득 시간을 알려주는 변수이다. 시간 변수라면 품질 데이터의 작업일 변수처럼 타입이 datetime으로 설정돼 있어야 하지만 TAG_MIN 변수는 object 타입으로 설정돼 있다. 따라서 TAG_MIN 변수를 datetime 타입으로 바꾸어 준다.

2. 데이터 결측치 처리

- 공정 데이터를 보면 상당 부분 결측값이 존재하는 것을 확인할 수 있다.

```
TAG_MIN                  0
배정번호                     0
건조 1존 OP                 1
건조 2존 OP                 1
건조로 온도 1 Zone          116
건조로 온도 2 Zone          148
세정기                     91
소입1존 OP                4288
소입2존 OP                   0
소입3존 OP                   2
소입4존 OP                   3
소입로 CP 값                  1
소입로 CP 모니터 값            147
소입로 온도 1 Zone          130
소입로 온도 2 Zone          128
소입로 온도 3 Zone          157
소입로 온도 4 Zone          170
솔트 컨베이어 온도 1 Zone      106
솔트 컨베이어 온도 2 Zone      142
솔트조 온도 1 Zone          209
솔트조 온도 2 Zone          203
```

<그림 4> 공정 데이터 결측치

변수명	상관계수
건조 1존 OP, 건조로 온도 1 Zone	-0.588599
건조 2존 OP, 건조로 온도 2 Zone	-0.437099
소입1존 OP, 소입로 온도 1 Zone	-0.759778
소입2존 OP, 소입로 온도 2 Zone	-0.499504
소입3존 OP, 소입로 온도 3 Zone	-0.592332
소입4존 OP, 소입로 온도 4 Zone	-0.718203

<표 3> OP, 온도간 상관계수표

- 각 존의 OP값이 온도값에 영향을 미칠 것이라 생각했고, 이를 확인하기 위해 각각의 변수간 상관관계를 확인하였다. 확인 결과 OP값이 온도에 영향을 미친다고 판단하여, 배정번호 내에서 선형회귀를 수행하여 각 결측값을 대치한다.

- 이후 남은 결측치들에 대해서는 배정번호별로 각 변수의 평균값으로 대치한다.

```
TAG_MIN                  0
배정번호                     0
건조 1존 OP                 1
건조 2존 OP                 1
건조로 온도 1 Zone            1
건조로 온도 2 Zone            1
세정기                     91
소입1존 OP                  26
소입2존 OP                   0
소입3존 OP                   1
소입4존 OP                   2
소입로 CP 값                  1
소입로 CP 모니터 값            147
소입로 온도 1 Zone           26
소입로 온도 2 Zone            0
소입로 온도 3 Zone            1
소입로 온도 4 Zone            2
솔트 컨베이어 온도 1 Zone      106
솔트 컨베이어 온도 2 Zone      142
솔트조 온도 1 Zone          209
솔트조 온도 2 Zone          203
```

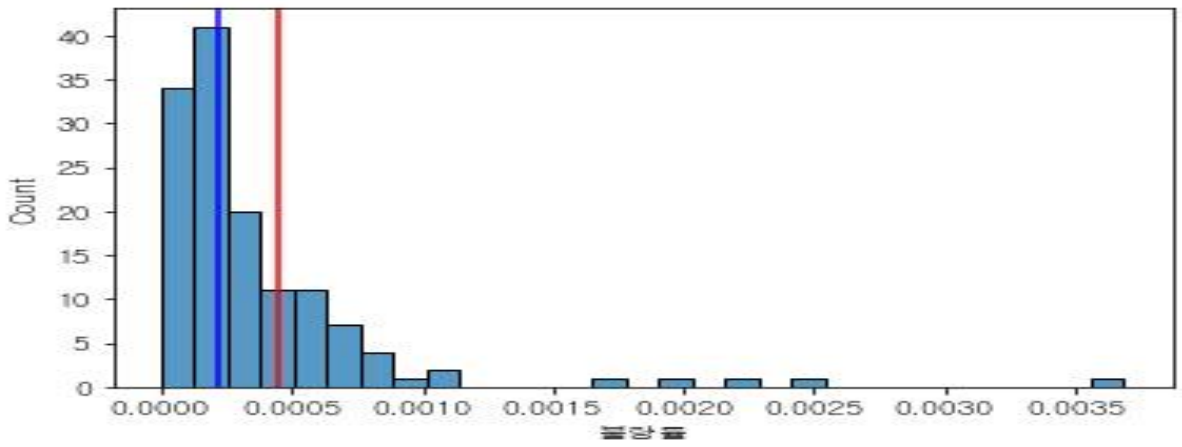
<그림 5> 선형회귀 수행 이후 공정 데이터 결측치

```
TAG_MIN                  0
배정번호                     0
건조 1존 OP                 0
건조 2존 OP                 0
건조로 온도 1 Zone            0
건조로 온도 2 Zone            0
세정기                      0
소입1존 OP                  0
소입2존 OP                  0
소입3존 OP                  0
소입4존 OP                  0
소입로 CP 값                  0
소입로 CP 모니터 값            0
소입로 온도 1 Zone            0
소입로 온도 2 Zone            0
소입로 온도 3 Zone            0
소입로 온도 4 Zone            0
솔트 컨베이어 온도 1 Zone      0
솔트 컨베이어 온도 2 Zone      0
솔트조 온도 1 Zone            0
솔트조 온도 2 Zone            0
```

<그림 6> 평균값 대치 이후 공정 데이터 결측치

3. 데이터 병합/변수 추가 및 제거

- 결측치 대치 이후 상관계수를 확인하여 '소입로 CP 모니터 값', '솔트조 온도 1 Zone'의 상관계수가 다른 변수보다 큰 것을 확인하였다. 상관관계가 클 경우 다중공선성 문제가 발생해 분석에 영향을 미칠 수 있다. 따라서 상관관계가 높은 두 변수를 제거한다.
- 공정 품질을 예측하기 위해서는 공정 데이터와 품질 데이터 두 개를 활용해야 하므로 두 데이터를 병합할 필요가 있다. 하지만 지금 형태로는 두 데이터를 합칠 수 없기 때문에 공정 데이터를 변형시켜야 한다. groupby() 함수를 사용하여 배정번호별로 그룹핑 한 이후 각 배정번호별 변수들의 평균, 표준편차, 최대값, 최소값, 제1사분위수, 중위수, 제3사분위수, 변동계수를 추가한 데이터셋을 만든다.
- 새롭게 만든 공정 데이터셋과 품질 데이터셋을 합쳐주기 전에, 품질 데이터셋에서 필요없는 변수들을 제거한다. 작업일 변수의 경우 TAG_MIN 변수가 이미 있으므로 제거하고, 공정명 변수와 설비명 변수의 경우 값이 1개밖에 없어 분석에 사용하지 않는 변수이므로 제거한다.
- 양품수량, 불량수량, 총수량 변수의 경우 새로운 변수인 불량률을 추가하는데 사용하고, 사용한 이후에는 분석에 사용하지 않는 변수이므로 제거한다. 불량률 변수의 경우 불량률 = 불량수량 / 총수량 으로 계산하여 구한다.
- 병합 이후 불량률 변수로부터 불량단계 변수를 생성한다. 불량률이 기준 수치 이상을 넘어서면 해당 공정은'위험'으로 분류하고, 기준 수치 미만이면 해당 공정은'안정'으로 분류한다. 불량률의 분포를 살펴보면 오른쪽 꼬리가 긴 것을 확인할 수 있다. 따라서 분포의 중심 우측에 해당하는 기술통계량 중 하나인 3사분위수를 기준으로 불량단계를 구분하였다.



<그림 7> 불량률 변수의 히스토그램

4. 상관계수에 기반한 변수 제거

- 분석을 위해 만든 새로운 데이터셋은 많은 변수를 가지고 있다. 변수의 개수가 많을수록 높은 상관관계를 가지는 변수 또한 많아지는데, 상관계수가 높은 변수를 제거함으로써 모델의 단순화와 계산의 효율성을 증가시킬 수 있다. 따라서 상관계수 0.7을 기준으로 0.7보다 큰 값을 가지는 변수를 제거한다.
- 완성된 최종 데이터셋은 다음과 같다.

	배정번호	불량률	건조 1존 OP_Mean	건조 1존 OP_Std	건조 1존 OP_Max	건조 1존 OP_Min	건조 1존 OP_Q25	건조 1존 OP_Q50	건조 1존 OP_Q75	건조 1존 OP_CV	...	솔트 컨베이어 온도 2 Zone_Q75	솔트조 온도 2 Zone_Mean	솔트조 온도 2 Zone_Std	솔트조 도 Zone_M
0	102410	0.000198	72.252727	3.696537	83.4128	55.8907	71.119075	72.56910	73.919450	5.115773	...	285.9740	329.070466	0.116518	329.3
1	102585	0.000334	72.235643	3.365000	82.5948	55.4900	70.880000	72.54800	74.229450	4.658190	...	286.8560	328.924151	0.089118	329.1
2	102930	0.000503	70.720207	3.231776	81.7570	53.9100	69.656725	71.17550	72.591900	4.569722	...	285.5810	329.148656	0.115567	329.3
3	103142	0.000174	72.424229	2.635245	82.3991	55.6307	71.171650	72.54490	73.954300	3.638566	...	285.5720	329.073103	0.100487	329.2
4	103675	0.000134	72.774648	4.159221	84.5765	52.9816	71.879925	73.41015	74.829000	5.714746	...	282.2200	329.114051	0.078846	329.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
131	147292	0.000274	69.486127	3.123360	78.2255	53.7822	68.145100	69.90460	71.509500	4.494829	...	287.0540	332.210106	0.129175	332.4
132	147546	0.000267	69.718808	2.662344	79.5047	53.0555	68.678650	70.09770	71.344750	3.818619	...	286.4180	332.162465	0.260713	332.4
133	147982	0.000293	69.799029	3.164459	79.7104	54.4260	68.727000	70.32180	71.749950	4.533550	...	286.4390	332.125705	0.118017	332.3
134	147996	0.000298	69.991809	3.564122	80.8309	56.2065	68.682200	70.68730	72.386875	5.092040	...	287.2325	332.088618	0.237818	332.4
135	148069	0.000000	69.032823	3.781360	78.8542	50.2607	68.152050	69.76780	71.221050	5.477520	...	286.8420	332.319032	0.196851	332.6

136 rows × 107 columns

<그림 8> 분석에 활용할 데이터셋

5. 품질지수 측정내용

- 분석을 위해 완성한 최종 데이터셋의 품질 확인을 위해 품질지수를 구한다.
- 완전성 : 필수항목에 누락이 없어야 한다.
  최종 데이터셋의 완전성을 알아보기 위해 최종 데이터셋에 결측치가 있는지 없는지 확인한다.
- 유일성 : 데이터 항목은 유일해야 하며 중복되어서는 안 된다
  각 배정번호별로 groupby() 함수를 실행하여 최종 데이터셋을 만들었기 때문에 배정번호가 중복되어서는 안된다. 최종 데이터셋의 행 길이와 배정번호의 고유 데이터 개수가 같은지 확인한다.
- 유효성 : 데이터 항목은 정해진 유효범위 및 도메인을 충족해야 한다.
  공정 데이터, 품질 데이터, 최종 데이터의 배정번호 개수를 확인하여 개수가 서로 일치하는지 확인한다. 또한 각 배정번호별로 공정 데이터가 수집된 시간과 각 배정번호별 품질 데이터의 작업일이 일치하는지 확인한다.
- 일관성 : 데이터가 지켜야 할 구조, 값, 표현되는 형태가 일관되게 정의되고, 서로 일치해야 한다.
  최종 데이터의 각 변수들의 타입이 유효한지 확인한다.
- 정확성 : 실제 존재하는 객체의 표현 값이 정확하게 반영이 되어야 한다.
  품질 데이터의 양품수량, 불량수량, 총수량 변수를 확인하여 총수량 변수가 양품수량과 불량수량의 합이 정확히 맞는지 확인한다.
- 무결성 : 데이터베이스 자료의 오류 없이 변화에 영향을 받지 않고 데이터의 유일성, 유효성, 일관성이 보호되어야 한다.
  유일성, 유효성, 일관성이 모두 만족되면 데이터셋은 무결성을 만족한다고 할 수 있다

구분	품질지수
완전성	100%
유일성	100%
유효성	100%
일관성	100%
정확성	100%
무결성	100%

<표 4> 품질지수 측정

6. 모델 학습 데이터/검증 데이터 구분

- 주어진 데이터셋을 모델의 학습에만 사용하면 모델의 성능을 확인할 수 없고, 모델이 과적합 되었는지 확인할 수 있는 방법이 없다. 따라서 데이터셋을 학습 데이터와 검증 데이터로 나누어 모델의 성능과 모델의 과적합 여부를 확인한다.

7. SMOTE-Tomek

- 학습에 사용되는 데이터셋은 불량 단계가 '안정'인 데이터가 '위험'인 데이터보다 더 많아 균형잡힌 데이터가 아니다. 이러한 불균형은 모델의 성능을 저하시킬 수 있기 때문에 모델의 성능을 높이기 위해서는 데이터 불균형을 해결해야할 필요가 있다.
- SMOTE-Tomek은 오버샘플링(소수 데이터의 개수를 증가시키는 방법)과 언더샘플링(다수 데이터의 개수를 감소시키는 방법)을 함께 수행하는 방법으로, SMOTE 기법으로 오버샘플링을, Tomek Links로 언더샘플링을 수행하는 기법이다.
- SMOTE-Tomek을 사용함으로써 오버샘플링을 사용했을 때의 과적합 문제와 언더샘플링을 사용했을 때의 정보 손실 문제를 최소화 할 수 있다. 아래 표는 SMOTE-Tomek 사용 전과 후의 데이터셋의 분포인데, SMOTE-Tomek을 사용하여 데이터셋을 균형잡힌 데이터셋으로 만들 수 있었다.

SMOTE-Tomek 사용 이전 데이터셋의 분포		SMOTE-Tomek 사용 이후 데이터셋의 분포	
안정	70	안정	70
위험	25	위험	67

<표 5> SMOTE-Tomek 사용 이전/이후 데이터셋의 분포

□ 분석모델 개발

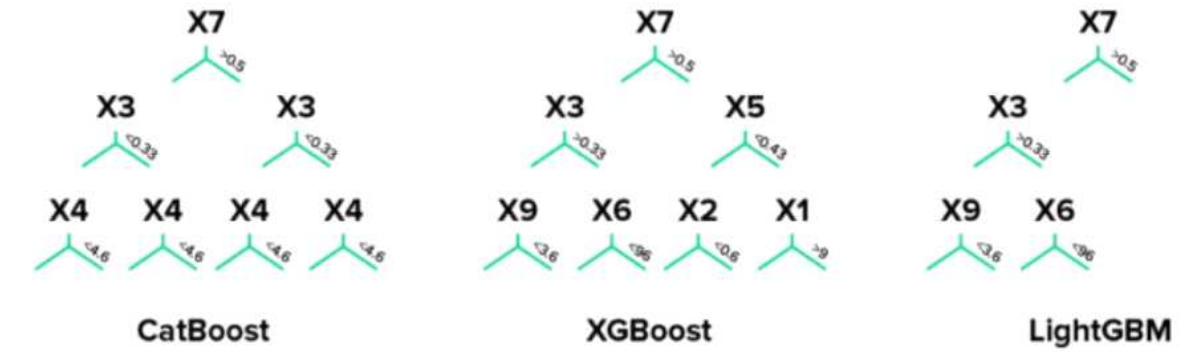
○ AI 분석 방법론(알고리즘)

- Catboost

- Catboost는 속도 개선 로직과 정규화 방법을 보유하고 있는 Boosting류 모델이다.

1. Level-wise Tree

- CatBoost도 XGBoost처럼 BFS 방식 즉, level-wise 방식으로 트리를 형성하나 Feature를 모두 일하게 대칭적인 트리 구조로 형성하게 된다. 이런 방식을 통해 예측 시간이 감소하게 된다.



<그림 9> CatBoost 트리 성장 예시

- CatBoost도 XGBoost처럼 BFS 방식 즉, level-wise 방식으로 트리를 형성하나 Feature를 모두 일하게 대칭적인 트리 구조로 형성하게 된다. 이런 방식을 통해 예측 시간이 감소하게 된다.

2. Ordered Boosting

- 기존의 부스팅 모델이 일괄적으로 모든 훈련 데이터를 대상으로 잔차계산을 했다면, CatBoost는 일부 데이터만을 가지고 잔차계산을 한 뒤, 모델을 만들어 나머지 데이터의 잔차는 이 모델로 예측한 값을 사용한다. 이는 Boosting모델에 오버피팅을 방지해주는 역할을 한다.

- Catboost의 장점

- 빠른 처리속도: Catboost는 효율적인 알고리즘을 구현하여 상대적으로 빠른 학습 속도를 제공한다.
- 불균형 데이터 처리 및 과적합 방지: 불균형 데이터셋도 파라미터 조정을 통해 예측력을 높일 수 있다. 또한 과적합을 피하기 위해 내부적으로 여러 방법(random permutation, overfitting detector)을 갖추고 있어 성능 또한 뛰어나다.

- XAI(eXplainable Artificial Intelligence)

- 설명 가능한 인공지능(XAI)은 인공지능에 의해 예측된 결과를 설명하여, 사용자가 결과를 이해할 수 있도록 돕는 것을 말한다. 이는 결과에 미치는 주요 요인들을 찾아내어 기계 학습 모델의 예측 결과를 어떤 근거로 의사 결정을 내렸는지를 알 수 있게 하며, 예측 결과에 대해 사람이 이해할 수 있는 직관적인 설명을 가능하게 한다.

- SHAP(Shapley value)

- SHAP는 학습 데이터와 학습된 모델을 바탕으로 설명 가능한 모델을 생성하고 새로 입력된 데이터에 대해 예측 결과에 대한 영향력을 방향과 크기로 표현한 Shapley Values을 계산한다. 이를 통해, 입력 변수가 학습된 모델의 출력값에 어느 정도의 공헌도를 가지는지 설명한다.

- SHAP의 장점

- 기존의 변수 중요도(Feature Importance) 기법은 순열방법을 사용해서 변수가 모델에 미치는 영향을 측정한다. 이 방법은 계산 속도가 빠르다는 장점이 있지만, 변수들이 서로 의존적일 때는 결과가 왜곡될 수 있다. 또한, 음(-)의 영향력은 계산하지 못한다. 따라서 실제 영향력보다 특정 변수의 가치가 높게 책정될 수 있다.
- 반면에 SHAP 기법은 변수들이 서로 영향을 미칠 가능성을 고려하고 음(-)의 영향력을 계산할 수 있다. 그래서 속도가 느리다는 단점이 있지만, 변수 중요도 기법보다 정확한 영향력을 측정한다고 볼 수 있다.

- Confusion Matrix

- 분류 모델에 대한 성능 평가 지표로 사용되며 Training을 통한 Prediction 성능을 측정하기 위해 예측 value와 실제 value를 비교하기 위한 표이다.

□ 분석결과 및 시사점

○ 분석 결과

- Confusion Matrix

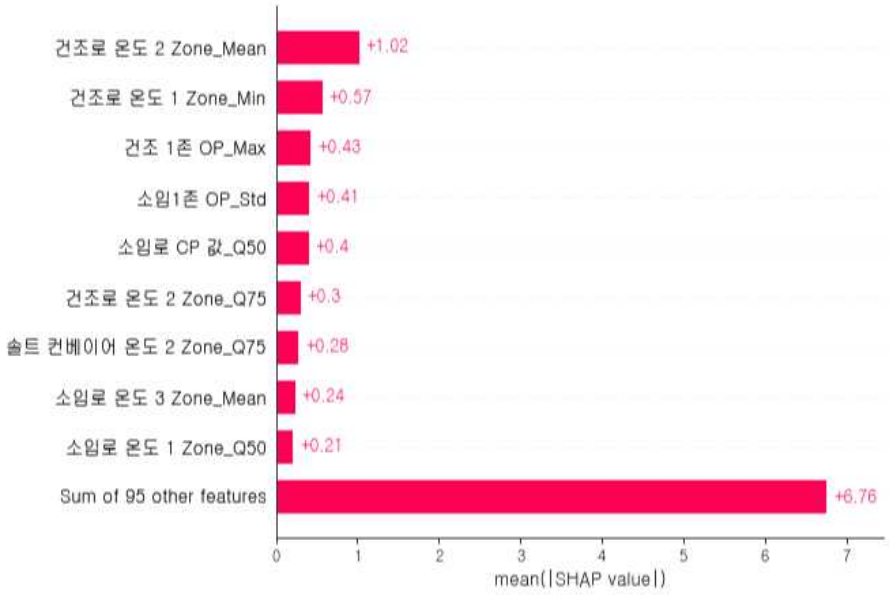
	precision	recall	f1-score	support
안정	0.85	0.88	0.86	32
위험	0.5	0.44	0.47	9
Accuracy			0.78	41
Macro avg	0.67	0.66	0.67	41
Weighted avg	0.77	0.78	0.78	41

<표 6> Confusion Matrix

- 분석 결과는 위 그림과 같다. 해당 모델의 경우‘안정’을 예측하는 f1-score 값은 0.86이고, ‘위험’을 예측하는 f1-score 값은 0.47이다. ‘위험’을 예측하는 f1-score 값이 좋지 않은데, 하이퍼파라미터 튜닝이나 변수 선택 등의 기법을 효과적으로 활용하면 더욱 성능을 끌어올릴 수 있을 것이다.

– Shap Value

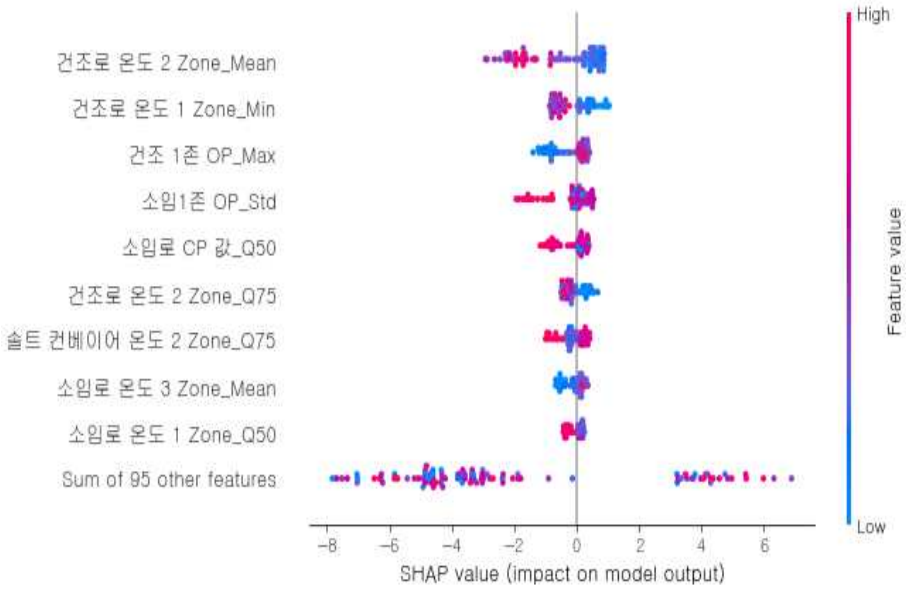
1. shap.plots.bar



<그림 10> 변수별 Shap 값의 절대 영향도 시각화

- Shap을 통한 각 변수가 모델에 미치는 절대 영향도는 다음 그림과 같다. 불량 단계를 예측하는 데 가장 큰 영향을 미치는 변수는'건조로 온도 2 Zone_Mean'이며 다음으로는 '건조로 온도 1 Zone_Min', '건조 1존 OP_Max'이다.
- 건조로와 소입로에 관련된 변수들이 불량 단계를 예측하는 데 많은 영향을 미치는 것을 확인할 수 있다.

2. shap.plots.beeswarm

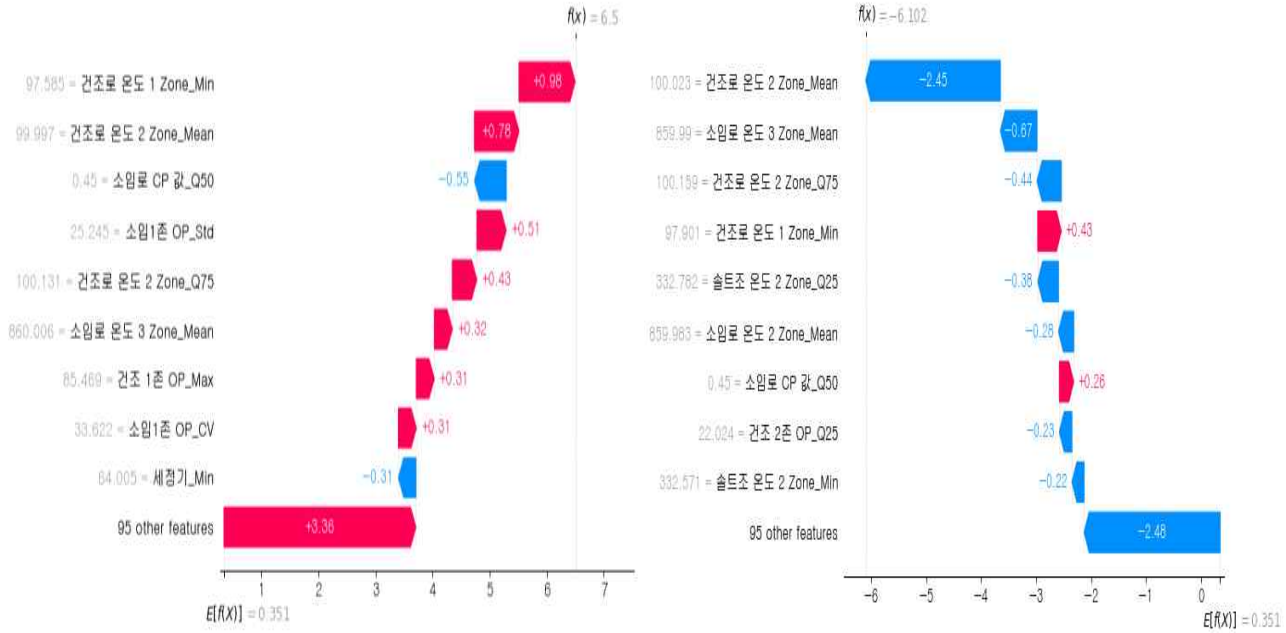


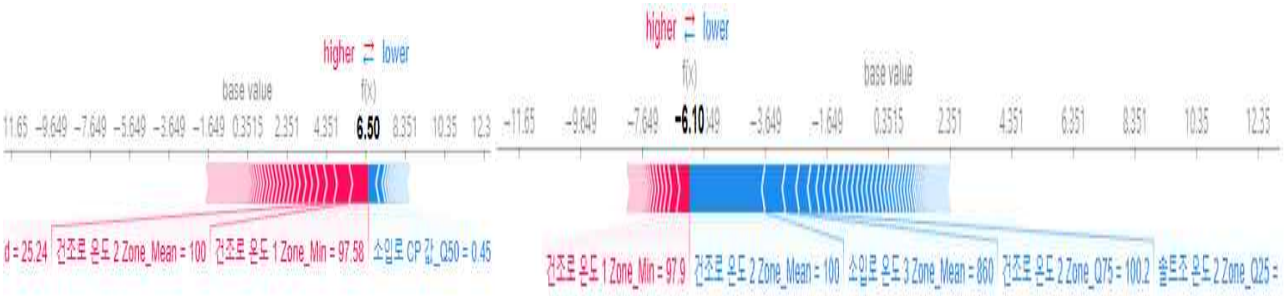
<그림 11> Shap 요약 그래프로

- 위 그림은 Shap 요약 그래프로, 각 변수가 불량 단계 예측에 미치는 영향 정도를 나타내는 그림이다.

- 빨간색에 가까운 점은 변수가 큰 값을 의미하고, 파란색에 가까운 점은 변수가 작은 값을 의미한다. 또한 점이 왼쪽으로 치우쳤을 경우 불량 단계를 '안정'이라고 예측할 확률이 높다는 걸 의미하고, 점이 오른쪽으로 치우쳤을 경우 불량 단계를 '위험'이라고 예측할 확률이 높다는 걸 의미한다.
- 예를 들어, '건조로 온도 2 Zone_Mean'의 경우 값이 클수록 불량 단계를 안정이라고 예측할 확률이 높다는 것을 의미하고 값이 작을수록 불량 단계를 위험이라고 예측할 확률이 높다는 것을 의미한다.

3. shap.plots.waterfall/shap.plots.force





- 오른쪽 그림은 배정번호 113977 데이터로 얻은 결과 예시이며, 왼쪽 그림은 배정번호 133605 데이터로 얻은 결과 예시이다.
- 해당 변수들이 예측 결과에 어느정도 크기 및 방향으로 영향을 주었는지 알 수 있다. 붉은색 변수는 불량 단계를 위험이라고 예측하는데 영향을 준 변수이고, 파란색 변수는 불량 단계를 안정이라고 예측하는데 영향을 준 변수이다.
- 배정번호 113977 데이터가 불량이라고 예측되는 데 영향을 준 변수는 '건조로 온도 1 Zone_Min'이 가장 크다고 해석할 수 있다. 또한 배정번호 133605 데이터가 안정이라고 예측되는 데 영향을 준 변수는 '건조로 온도 2 Zone_Mean'이 가장 크다고 해석할 수 있다.

○ 시사점 및 보완점

- 시사점

- shap.plots.bar와 shap.plots.beeswarm의 변수 영향도 결과를 종합해본 결과 건조로와 소입로가 불량 단계를 예측하는데 큰 영향을 주는 것을 확인할 수 있었다. 따라서 건조로와 소입로에 대한 관리에 따라 불량 단계가 결정된다고 할 수 있다.
- 건조로와 소입로에 대한 관리를 강화하여 건조로와 소입로를 효과적으로 제어할 수 있다면 불량률을 안정 단계 수준으로 관리할 수 있을 것이다.

– 보완점

- 열처리 공정에서 사용되는 다른 변수들을 사용할 수 있다면 더 좋은 결과를 얻을 수 있을 것이다.
- 데이터를 샘플링하는 과정에서 더 많은 데이터가 있다면 모델의 예측 성능을 더욱 끌어올릴 수 있을 것이며, 더 좋은 결과를 얻을 수 있을 것이다.
- 불량 단계를 구분할 수 있는 더욱 확실한 기준이 있다면 수준 높은 분석 결과를 얻을 수 있을 것이다.

□ 중소제조기업에 미치는 파급효과

○ 파급효과

- 공정 데이터를 이용하여 생산품의 품질을 예측할 수 있다면 공정 운용에 큰 도움이 될 것이다. 위와 같은 방법을 적용하면 생산품의 품질을 결정하는데 중요한 요소가 무엇인지 파악할 수 있으며 그에 대한 대책을 마련할 수 있을 것이다.

- 또한 분석의 전체적인 프로세스를 제공했기 때문에 위와 같은 전처리가 가능하다면 타 공정 및 타 분야로도 확장이 가능할 것으로 보인다.sS