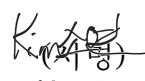
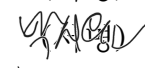



제3회 K-인공지능 제조데이터 분석 경진대회 보고서

프로젝트명	열처리 공정 품질 예측을 위한 가중 평균 앙상블 모델 개발
팀명	BAA
내용요약	<p>열처리 공정에서 실시간 품질 확인의 어려움을 해결하기 위해 실시간 데이터를 활용하여 주요 인자를 추출하고 품질을 사전에 예측하는 품질보증 모델을 개발하여 공정을 효과적으로 제어하고 모니터링하는 방법을 제안한다.</p> <p>먼저, 컬럼 간의 상관관계를 분석하여 높은 상관관계를 가지는 컬럼을 처리하여 다중공선성을 방지한다. 이렇게 함으로써 데이터의 중복성을 줄이고 모델의 정확도를 향상시킬 수 있다.</p> <p>수치 데이터는 불량률을 기준으로 '위험'과 '안정'으로 오토라벨링한다. 이렇게 오토라벨링된 데이터를 활용하여 효율적인 데이터 분석과 모델링을 수행한다.</p> <p>XGBoost와 중요도 점수 계산을 사용하여 특성을 평가하고, RFE(Recursive Feature Elimination)를 활용하여 중요도가 낮은 특성을 순차적으로 제거하여 목표 특성 개수에 도달한다. 이 단계는 모델의 특성을 최적화하는데 도움을 준다.</p> <p>XGBoost와 신경망 모델을 결합하여 모델의 성능을 향상시키며, XGBoost 모델이 뛰어난 성능을 보이므로 해당 모델에 높은 가중치를 부여합니다. 두 모델을 결합하여 예측 성능이 우수한 가중 평균 앙상블 모델을 개발하였으며, 해당 모델의 F1-Score는 0.9630이고 정확도는 0.9412이다.</p> <p>이 분석 모델을 활용하면 공정 조건에 따른 품질 예측을 통해 문제를 예방할 수 있다. 이는 원가를 절감하고 품질을 향상시키며, 자동화와 효율성 향상을 이룰 수 있다. 또한, 이를 통해 고객 만족도를 높이고 경쟁력을 강화할 수 있을 것으로 기대된다.</p>
<p>상기 본인(팀)은 위의 내용과 같이 제3회 K-인공지능 제조데이터 분석 경진대회 결과 보고서를 제출합니다.</p> <p>2023 년 11월 3일</p> <p>팀장 : 김소연 </p> <p>팀원 : 박시운 </p> <p>팀원 : 최예진 </p> <p>한국과학기술원장 귀중</p>	

□ 문제정의

○ 분석 내용

- 공정 개요

금속 열처리란 금속의 성질을 개선하기 위해 금속 표면에 가열과 냉각을 조절하여 행하는 기술이다. 금속 재료의 가공 이후에는 원하는 물성을 얻기 위한 열처리 공정이 필수적으로 진행된다. 금속 열처리의 종류로는 크게 강화, 연화, 잔류 응력 제거가 있다. 이러한 금속 열처리에 같은 조건으로 공정을 적용해도 최종 제품의 물성이 달라지는 문제가 있기 때문에 물성이 달라지는 원인을 파악하면 제품의 수율을 안정시킬 수 있다.

열처리 방법은 다양한 종류로 나눌 수 있는데, 주요 방법으로는 어닐링(Annealing)을 통해 결정립을 미세화하여 기계적 성질이나 가공성을 향상시키는 방법, 노멀라이징(Normalizing)을 통해 부품을 균질화시키는 방법, 퀴칭(Quenching)을 활용하여 부품을 경화하는 방법, 템퍼링(Tempering)을 통해 부품을 강인화하는 방법, 그리고 과냉 오스테나이트가 변태 완료할 때까지 염욕에 담금질하여 항온을 유지한 후 공랭하는 오스템퍼링(Austempering) 방법이 있다.

해당 분석 데이터의 열처리 공정은 전기 가열식 설비를 사용하며, 소입, 염욕, 세척, 건조 단계를 거친다. 먼저, 소입로에서 가스 침탄과 오스테나이트화가 이루어지고 염욕 단계에서는 250℃에서 450℃의 온도 범위에서 작업을 진행하여 베이나이트 조직을 얻는다. 이 과정은 주로 부품의 경화와 인성을 향상시키기 위해 수행되며, 이렇게 함으로써 원하는 부품 특성을 얻는 것이 해당 열처리 공정의 목표이다.

- 분석 필요성

열처리 공정에서 발생하는 주요 문제는 균열, 변형, 얼룩, 조직 불량 등이 있다. 재료 품질 향상, 설비 개선, 온도 제어 기술 발전 등 다양한 개선 노력이 있었지만, 여전히 문제점이 남아있다.

특히, 열처리 과정에서의 문제는 생산품 투입부터 배출까지의 소요 시간이 약 2시간 이상이며, 불량을 눈으로 확인하기 전까지 제품 품질을 확인하기 어렵다는 것이다. 또한, 불량 발생 시 정확한 원인 파악이 어려워 개선 활동에 어려움을 겪고 있다.

이러한 어려움을 극복하기 위해 데이터 분석 기반의 품질 예측 방법을 통해 제품 품질을 예측하고 공정을 조절함으로써 열처리 공정의 문제점을 효과적으로 극복할 수 있다.

○ 분석 목적

공정 중 생산품의 상태를 직접 확인할 수 없는 상황에서 공정 품질을 확보하기 위해서는 데이터 기반의 품질 예측이 필수적이다. 실시간으로 변화하는 공정 데이터에서 영향을 미치는 주요 인자들을 추출하고 모델링하여 생산품의 품질을 사전에 예측하면 공정을 효과적으로 제어할 수 있다.

본 분석에서는 열처리 공정 데이터에서 주요 인자를 추출하여 품질을 예측하고 공정을 효과적으로 제어할 수 있는 품질보증 모델 구축을 목표로 한다. 이 분석 모델을 통해 열처리 공정의 품질을 사전에 예측하고 사전에 조치를 취할 수 있을 뿐만 아니라 공정 모니터링도 가능해질 것이다.

□ 제조데이터 정의 및 처리과정

○ 제조데이터 정의

- 데이터 유형 및 구조

- 제조 데이터 분야 및 공정명 : 자동차 부품(에어백, 안전벨트)의 열처리 공정 데이터
- 데이터 수집 기간 : 2022년 01월 ~ 2022년 07월
- 데이터 구조
 - data.csv : 21열 2,939,722행 61,734,162데이터
 - quality.xlsx : 7열 136행 952데이터

- 데이터 속성 및 타입 정의

- 원본 공정 데이터

속성	설명	타입
TAG_MIN	데이터 수집 시간	object
배정번호	공정의 작업 지시 번호	int
건조 1~2존 OP	각 건조 온도 유지를 위한 출력량(%)	float
건조로 온도 1~2 Zone	각 건조로 Zone의 온도 값	float
세정기	세정기 온도 값	float
소입1~4존 OP	각 소입존 온도 유지를 위한 출력량(%)	float
소입로 CP 값	침탄 가스의 침탄 능력의 양(%)	float
소입로 온도1~4 Zone	각 소입로 Zone의 온도 값	float
솔트 1~2존 OP	각 솔트존 온도 유지를 위한 출력량(%)	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 정의

- 원본 품질 데이터

속성	설명	타입
배정번호	공정의 작업 지시 번호	int
작업일	공정 작업 날짜	datetime
공정명	공정의 이름	object
설비명	설비의 이름	object
양품수량	양품의 생산 수량	int
불량수량	불량품의 생산 수량	int
총수량	전체 제품의 생산 수량	int

[표 2] 품질 데이터 정의

○ 제조데이터 전처리 및 EDA 과정

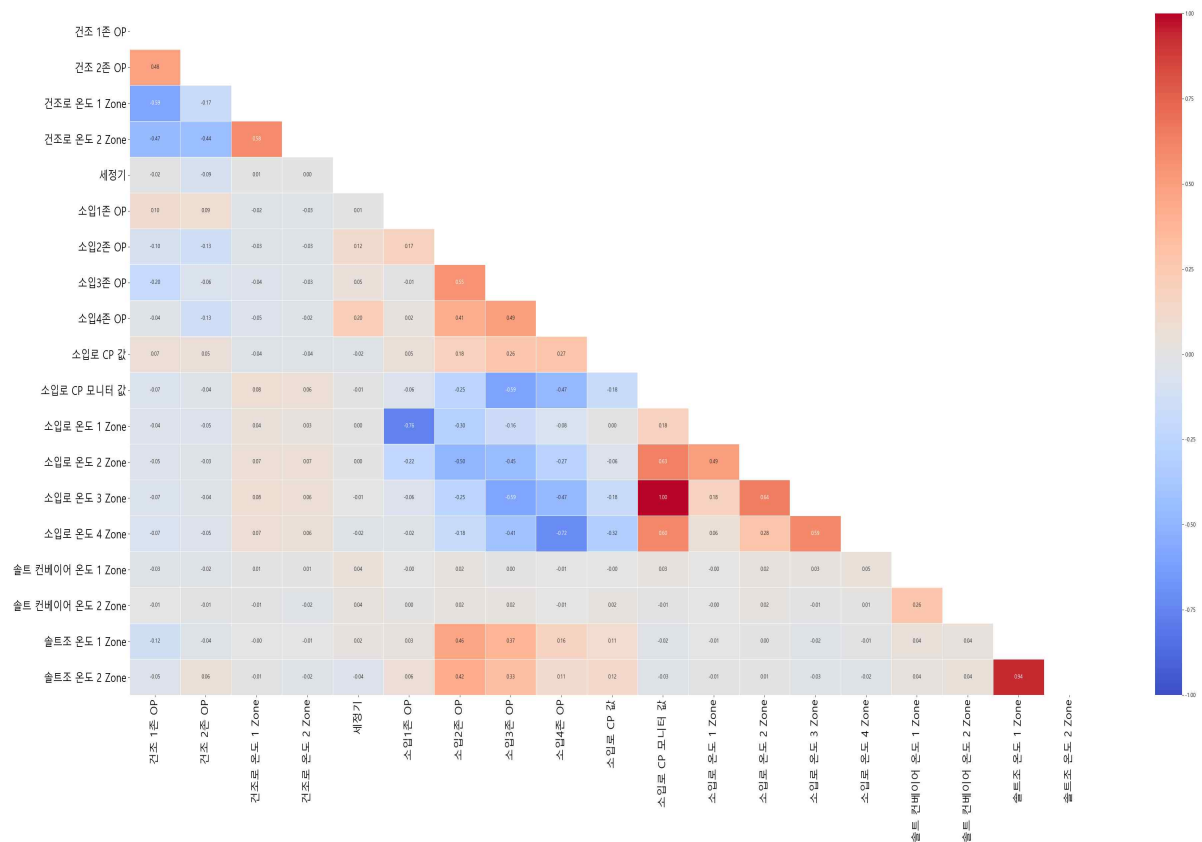
- 결측치 및 중복 데이터 처리

각 컬럼의 결측치를 확인한 결과, 'TAG_MIN', '배정번호', '소입2존 OP'을 제외한 모든 컬럼에 결측치가 존재함을 확인했다. 배정번호별로 데이터를 그룹화하여 결측치 비율을 확인한 결과, 최대 결측 비율은 0.087로 매우 낮았기 때문에 결측치를 제거하기로 결정했다.

또한, 행을 기준으로 중복된 데이터를 확인한 결과 중복이 없음을 확인했다. 'TAG_MIN' 컬럼을 제외한 상태에서 중복을 확인하면 중복 데이터가 존재했지만, 시계열 데이터의 특성상 이러한 데이터를 제거할 필요가 없다고 판단했다.

- 상관관계 확인

'TAG_MIN'과 '배정번호' 컬럼을 제외한 상관관계 그래프를 그려보았으며 그래프는 다음과 같다.

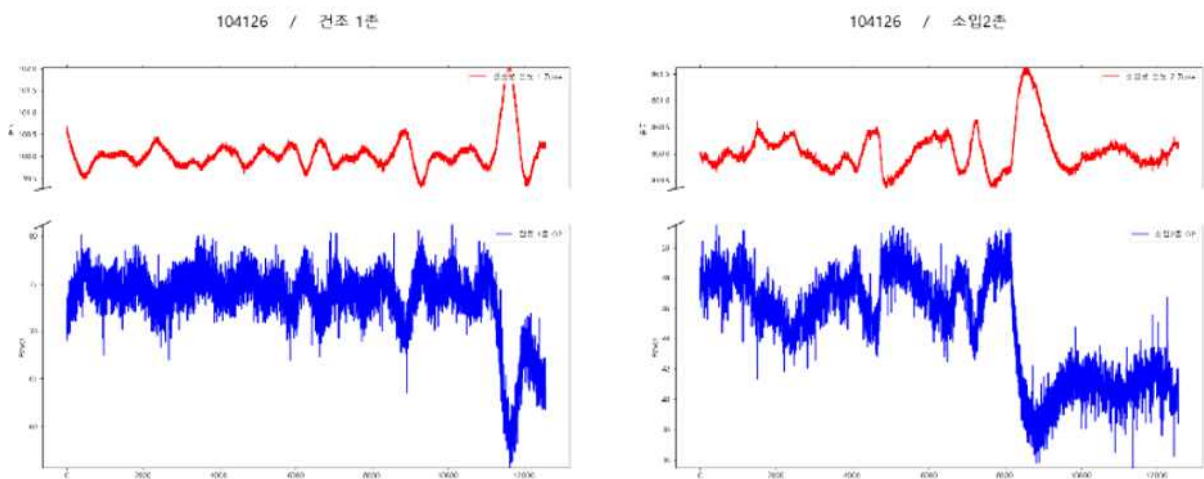


[그림 1] 컬럼별 상관관계 그래프

컬럼 간의 상관관계를 확인한 결과, 일부 컬럼 간에 높은 상관관계가 나타났다. 이러한 높은 상관관계는 다중공선성 문제를 발생시킬 수 있으며, 이로 인해 모델이 불안정해지고 변수의 중요도가 왜곡될 수 있다. 따라서 상관관계 확인을 통해 다중공선성을 발생시킬 가능성이 있는 컬럼을 사전에 처리하기로 했다.

OP 컬럼명	온도 컬럼명	상관계수
건조 1존 OP	건조로 온도 1 Zone	-0.6
건조 2존 OP	건조로 온도 2 Zone	-0.4
소입1존 OP	소입로 온도 1 Zone	-0.8
소입2존 OP	소입로 온도 2 Zone	-0.5
소입3존 OP	소입로 온도 3 Zone	-0.6
소입4존 OP	소입로 온도 4 Zone	-0.7

[표 3] 높은 음의 상관관계를 가지는 컬럼



[그림 2] 컬럼별 OP-온도 시각화

먼저, OP 컬럼과 온도 컬럼 사이에 높은 음의 상관관계가 있음을 확인했다. 각 OP 값은 온도를 유지하기 위한 출력량으로, 각 Zone의 온도와 관계가 있는 것을 [그림 2]의 '컬럼별 OP-온도 시각화'로 확인했다. 열처리 공정에서 각 Zone의 온도가 더 중요하다고 판단하여, 각 Zone의 OP 컬럼 6개를 제거하기로 결정했다.

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

[표 4] 매우 높은 양의 상관관계를 가지는 컬럼

그 외에도 높은 상관관계를 가지는 컬럼들을 확인했다. '소입로 CP 모니터값'과 '소입로 온도 3 Zone'간에는 매우 높은 상관관계가 나타났다. '소입로 CP 모니터값' 컬럼은 '소입로 온도 3 Zone' 이외의 컬럼들과도 높은 상관관계를 가지고 있었으며, 모니터값보다 온도가 열처리 공정에 더 큰 영향을 미칠 것이라고 판단하여 모니터값 컬럼을 제거하기로 결정했다. 그러나 '솔트조 온도 1 Zone'과 '솔트조 온도 2 Zone'은 온도 값을 나타내는 변수이므로 열처리 공정에 영향을 미칠 것이라고 가정하여, 이후의 변수 선택 모델을 통해 최종 결정하기로 했다.

- 통계량 컬럼 추가 및 데이터 통합

공정 데이터와 품질 데이터를 통합하기 위해서는 두 데이터의 차원을 맞추는 작업이 필요하다. 본 분석에서는 열처리 공정의 특성상 재료의 품질을 일정하게 유지하는 것이 중요하다고 판단하여, 각 배정번호별 변수의 통계량 컬럼을 추가하여 데이터를 통합하기로 결정했다. 공정 데이터를 '배정번호' 컬럼을 기준으로 그룹화한 후, 변수들의 평균값, 표준편차, 최솟값, 최댓값을 계산하여 새로운 컬럼으로 추가했다. 그 후, 배정번호를 기준으로 공정 데이터와 품질 데이터를 결합했다.

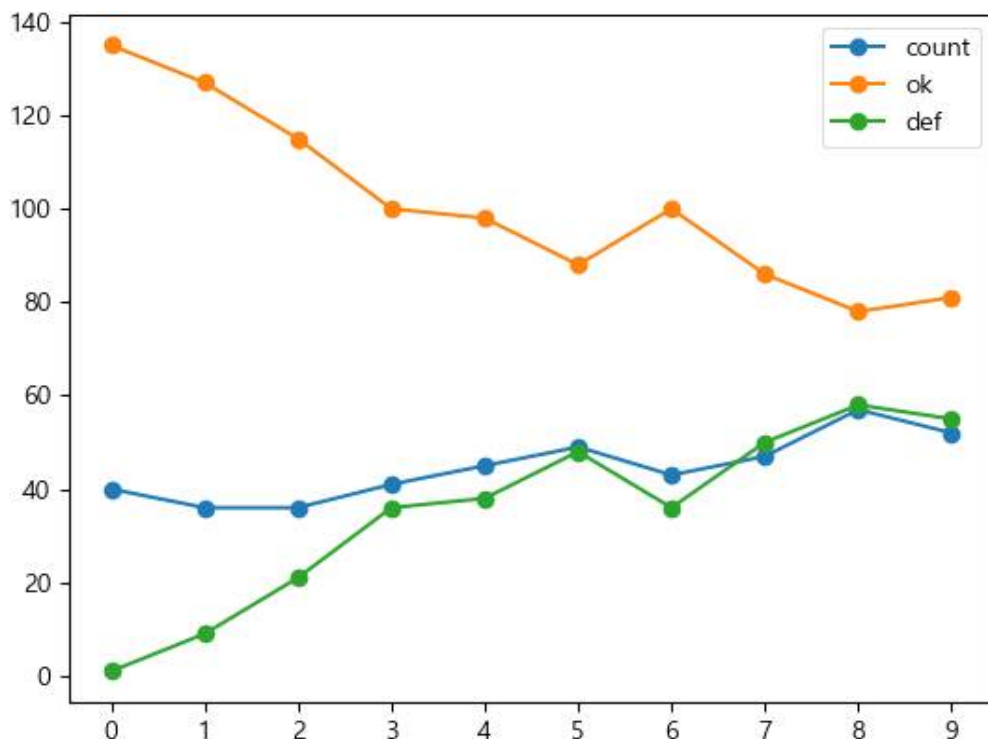
평균값은 대량의 데이터를 간결하게 요약할 수 있는 대푯값으로 사용했으며, 표준편차는 공정 중 발생하는 변동이 품질에 영향을 미칠 것으로 판단하여 사용했다. 최솟값과 최댓값은 공정 중 발생할 수 있는 극한 조건을 파악하여 품질을 예측하는데 활용하기 위해 사용했다.

이후, 통합한 데이터의 '불량수량' 컬럼을 '총수량' 컬럼을 나눈 후 100을 곱하여 '불량률' 컬럼을 새롭게 추가했다.

- Auto Labeling

수치 데이터에 대한 오토라벨링은 머신러닝 알고리즘 및 통계적 기법을 활용하여 레이블을 할당하거나 생성하는 프로세스이다. 이 프로세스는 데이터의 특성과 패턴을 이해하고 분류, 군집화, 회귀 및 이상치 탐지를 통해 데이터를 분석하여 레이블을 자동으로 부여한다. 수치 데이터 오토라벨링을 통해 데이터를 자동으로 레이블링 함으로써 효율적인 분석 및 모델링을 수행할 수 있다.

불량률을 기준으로 오토라벨링 기술을 활용하여 데이터 라벨링을 진행했다. 데이터 중에 불량률이 0%인 배정번호가 있어, 이 데이터를 사용하여 오토라벨링을 수행했다. 불량률을 기준으로, 상위 4개를 '위험(0)'으로, 하위 7개를 '안정(1)'으로 라벨링하는 모델로 최종 라벨링을 완료했다.



[그림 3] 양품과 불량품의 라벨 개수 그래프

개수의 기준은 다음과 같다.

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

기존과의 차이는 41개로 위험으로 분류한 경우는 36개, 안정으로 분류한 경우는 100개이다.

□ 분석모델 개발

○ 적용한 AI 분석 방법론

딥러닝은 복잡한 패턴 학습에 용이하지만 과적합의 위험이 있다. 반면, 머신러닝 모델은 더 단순한 패턴을 학습하므로 데이터의 작은 변화나 노이즈에 덜 민감할 수 있다. 딥러닝 모델이 특정 패턴이나 노이즈에 과적합되는 경우, 머신러닝 모델은 그러한 노이즈를 무시하고 주요 패턴에 집중한다. 따라서 두 모델이 서로 다른 오류 패턴을 가질 때, 두 모델을 조합하면 더 안정적인 예측이 가능하다.

- XGBoost - RFE

XGB-RFE는 XGBoost와 재귀적 특성 제거(RFE, Recursive Feature Elimination)를 조합한 특성 선택 방법이다. XGB-RFE 방법은 다음 절차를 따른다.

첫째, XGBoost를 사용한다. XGBoost는 각 특성의 중요도를 평가하는데 사용된다. 각 특성의 중요 포인트를 결정하고, 이를 기반으로 각 특성에 가중치를 부여한다.

둘째, 중요도 점수를 계산한다. 모든 부스팅 트리 내에서 각 특성의 가중치 합계를 사용하여 최종 중요도 점수를 계산한다.

셋째, 특성 순위를 설정한다. 계산된 중요도 점수에 따라 특성을 순위대로 정렬한다.

넷째, RFE를 사용한다. 중요도 순위를 기반으로 중요도가 낮은 특성을 순차적으로 제거한다. 이 과정을 목표 특성 개수에 도달할 때까지 반복한다.

XGB-RFE로 추출한 주요 컬럼
건조로 온도 1 Zone_min
건조로 온도 2 Zone_std
건조로 온도 3 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

[표 5] XGB-RFE로 추출한 10개의 주요 컬럼

XGB-RFE의 핵심 목적은 XGBoost의 특성 중요도 평가 능력과 RFE의 점진적 특성 제거 방법을 결합하여 모델 성능을 최적화하는 가장 중요한 특성들만 선택하는 것이다. 이 방법을 통해 불필요한 특성을 제거하여 모델의 복잡성을 줄이고 과적합을 방지하며 빠르고 효과적인 학습이 가능하다.

- XGBoost

XGBoost는 Chen과 Guestrin에 의해 소개된 분류모델로, XGBoost는 Gradient Tree Boosting 알고리즘에 확장이다. $n \times m$ 개의 데이터 세트 $D = (x_i, y_i) (|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$ 이고 K 개의 Tree를 사용하는 앙상블 모델이라고 할 때 수식은 다음과 같다.

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}, \quad (1)$$

여기서 $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}(q: \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ 는 회귀 트리이며 T 는 각 트리의 Leaf 개수이다. XGBoost는 이 트리들을 학습할 때 정규화된 손실 함수를 사용한다.

$$\begin{aligned} \mathcal{L}(\phi) &= \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \text{where } \Omega(f) &= \gamma T + \frac{1}{2} \lambda \|w\|^2 \end{aligned} \quad (2)$$

l 은 실제값과 예측값의 차이를 계산하는 미분 가능한 convex loss function이며 Ω 라는 정규화 텀을 두어 기존 Gradient Tree Boosting과 차별점을 둔다. 이 텀으로 T 가 작고 $\|w\|^2$ 이 작은 방향으로 학습한다. 따라서 모델이 과도하게 복잡해지는 것을 방지하고 과대적합을 예방하는데 도움을 준다.

XGBoost는 반복을 수행할 때마다 가지를 하나씩 늘려가는 방식을 사용한다.

$$\begin{aligned} \mathcal{L}^{(t)} &= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \\ \mathcal{L}^{(t)} &\simeq \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t) \\ \tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^n [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t) \end{aligned} \quad (3)$$

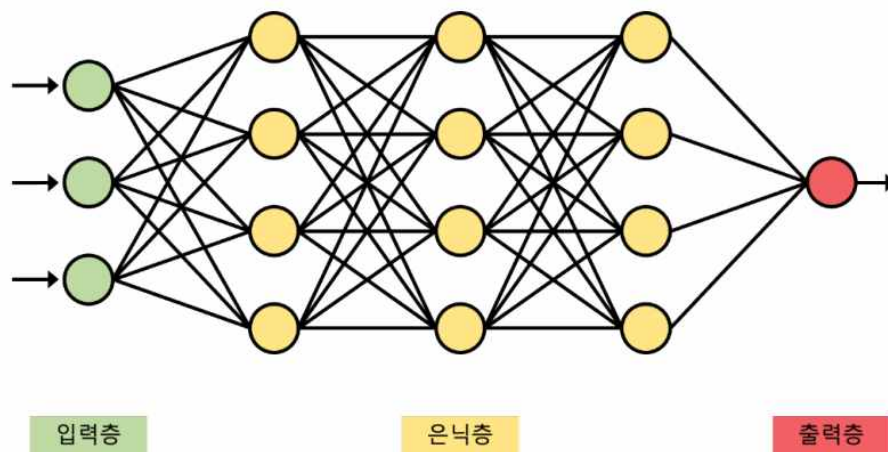
t 번째 손실 함수는 모든 데이터의 손실 함수를 더하고 정규화 텀을 추가한다. 직전 학습 단계에서의 예측값에 현재 예측값을 더해 손실 함수를 최소화한다. 따라서 앙상블 학습 과정에서 이전 학습에서 잘 학습하지 못한 부분에 가중치를 부여하고 학습한다.

본 분석에서는 RandomSearch 알고리즘으로 랜덤 샘플링 10회 수행하고, 교차 검증 5개의 폴드로 진행하여 가장 적합한 하이퍼파라미터를 결정했다.

훈련 정확도	0.9342105263157895
검증 정확도	0.8846153846153846

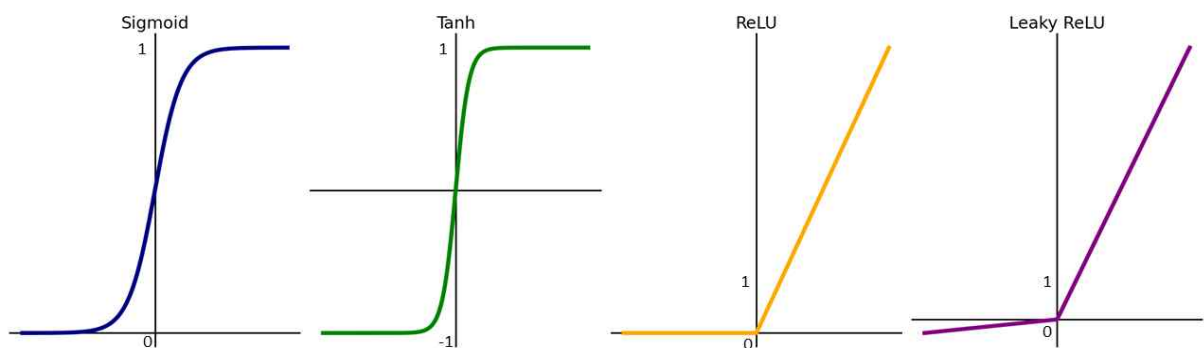
[표 6] 하이퍼파라미터 튜닝한 XGBoost 모델 정확도

- 인공신경망(ANN)



[그림 4] 인공신경망의 구조

신경망은 하나의 입력층, 하나 이상의 은닉층, 하나의 출력층으로 이루어져 있으며, 각 층은 여러 개의 노드로 구성된다. 은닉층에서는 행렬 곱과 같은 연속적인 연산을 수행한다. 활성화 함수(activation function)를 통해 선형이나 비선형 변환을 적용할 수 있으며, 은닉층을 여러 개 쌓으면 데이터의 복잡한 특성을 모델링할 수 있다.



[그림 5] 활성화 함수(activation function)의 종류

해당 분석에서 인공신경망을 이용한 이유는 XGBoost와 결합하여 예측 성능을 향상시키기 위함 이므로 층을 간단하게 쌓았다.

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	1408
dense_1 (Dense)	(None, 128)	16512
dropout (Dropout)	(None, 128)	0
batch_normalization (Batch Normalization)	(None, 128)	512
dense_2 (Dense)	(None, 1)	129
Total params: 18561 (72.50 KB)		
Trainable params: 18305 (71.50 KB)		
Non-trainable params: 256 (1.00 KB)		

[그림 6] 인공신경망 모델

첫 번째 'dense' 레이어는 입력층으로 128개의 뉴런을 가지며 입력 데이터의 특징과 연결된다. 활성화 함수로는 'relu'를 사용하여 비선형성을 도입하고 데이터의 복잡한 특성을 학습한다.

은닉층은 총 3개로 구성되어있다. 첫 번째 은닉층인 'dense_1' 레이어는 활성화 함수를 마찬가지로 'relu'를 사용했으며 데이터에서 더 복잡한 패턴을 학습할 수 있다. 그 다음 'Dropout' 레이어를 추가하여 50%의 뉴런을 무작위로 비활성화시켜 과적합을 방지한다. 그리고 'batch_normalization' 레이어를 통해 각 층의 입력을 정규화하여 학습 속도를 높이고, 초기화에 덜 민감하게 만든다. 이는 내부 공선성을 감소시켜서 최적화 문제를 안정화시키는데 도움을 준다.

마지막 'dense_2' 레이어는 출력층으로 사용되며 활성화 함수로 'sigmoid'를 사용하여 이진 분류를 수행한다.

훈련 정확도	0.9736841917037964
검증 정확도	0.9230769276618958

[표 7] 인공 신경망 모델 정확도

- XGBoost와 신경망 모델의 결합 (가중 평균 앙상블)

XGBoost와 신경망은 데이터를 각각 다르게 처리하고 학습하므로 데이터의 다양한 특성 파악이 가능하다. XGBoost는 트리 기반의 모델로 변수 간 상호작용을 잘 설명하며, 신경망은 계층적 구조로 복잡한 패턴과 비선형 관계를 파악하기에 좋다. 이러한 이유로 두 모델을 결합함으로써 각 알고리즘의 장점을 활용하여 모델의 성능을 향상시킬 수 있다. 그러나 여러 모델을 학습하고 튜닝해야 하기 때문에 전체 시스템의 복잡성이 증가할 수 있고, 이로 인해 계산 비용이 증가할 수 있다.

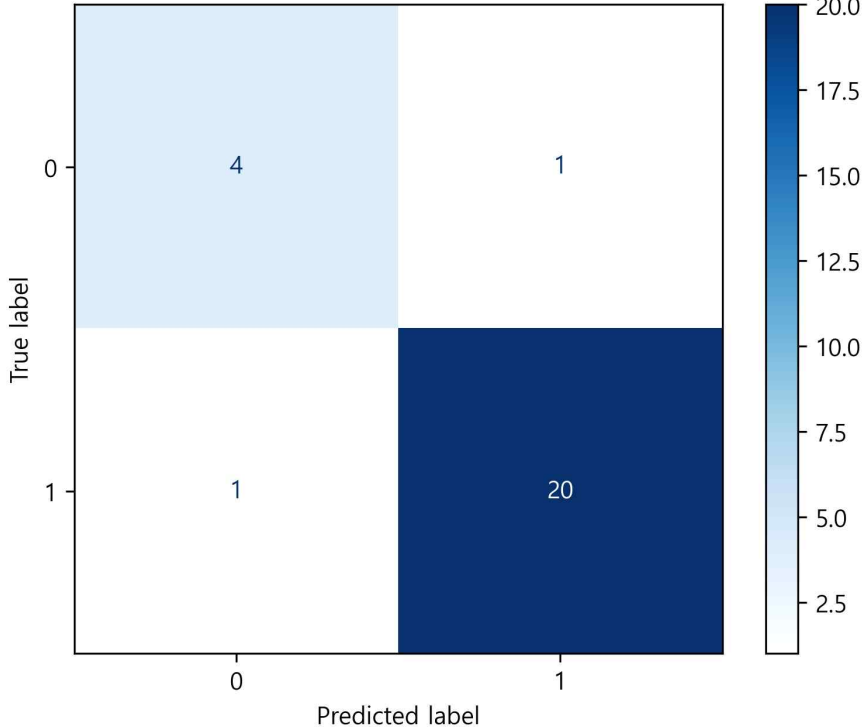
$$final\ prediction = \alpha \times prediction_{model} + (1 - \alpha) \times prediction_{model}$$

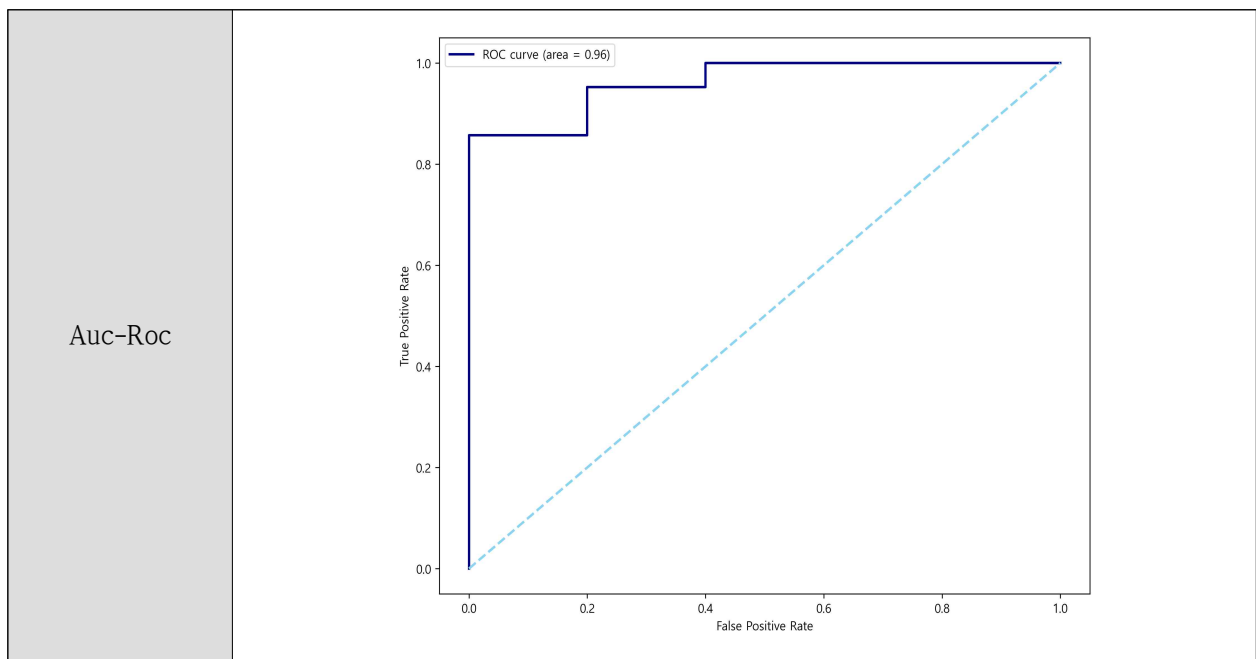
여기서 가중치 α 는 0과 1 사이의 값으로 각 모델의 신뢰도나 성능을 반영한다. 한 모델이 다른 모델보다 성능이 뛰어나다면 그 모델에 더 높은 가중치를 할당하여 최종 모델을 만들어야 한다.

본 분석에서는 XGBoost와 인공신경망 모델에 각각 0.6과 0.4의 가중치를 적용했다. 인공신경망의 훈련 및 검증 정확도는 XGBoost보다 상대적으로 높았지만, 데이터의 수가 제한적이어서 신경망의 결과가 초기 가중치나 데이터 변화에 민감할 것으로 예상되었다. 이러한 이유로 최종 모델은 신경망에 더 낮은 가중치를 부여했고 검증 데이터를 기준으로 0.65의 임계값을 설정하여 구성했다.

□ 분석결과 및 시사점

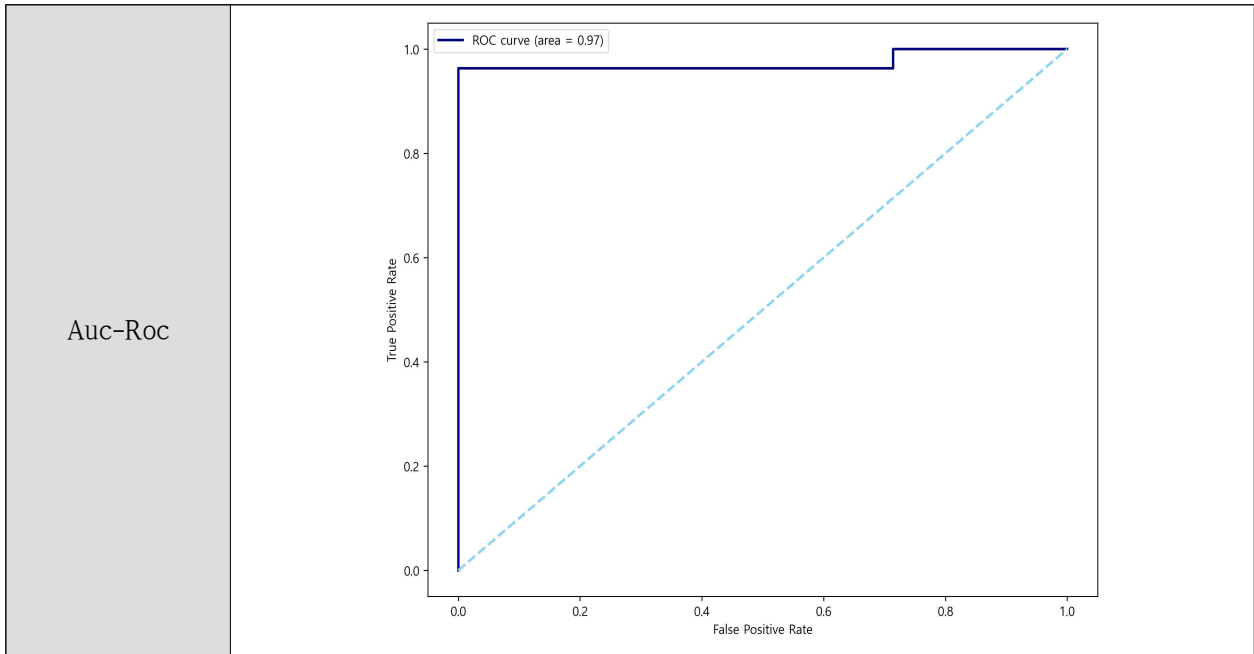
○ 분석 모델 성능 지표

구 분	검증 데이터
Accuracy	0.9231
Recall	0.9524
F1 Score	0.9524
AUC value	0.95619
Confusion Matrix	 <p>The confusion matrix is a 2x2 grid. The y-axis is labeled 'True label' with values 0 and 1. The x-axis is labeled 'Predicted label' with values 0 and 1. The cells contain the following counts: True label 0, Predicted label 0 is 4; True label 0, Predicted label 1 is 1; True label 1, Predicted label 0 is 1; True label 1, Predicted label 1 is 20. A color bar on the right shows a gradient from light blue (low values) to dark blue (high values), with tick marks at 2.5, 5.0, 7.5, 10.0, 12.5, 15.0, 17.5, and 20.0.</p>



[표 8] 최종 분석 모델 검증 성능 지표

구 분	테스트 데이터
Accuracy	0.9412
Recall	0.9630
F1 Score	0.9630
AUC value	0.9735
Confusion Matrix	<p>True label</p> <p>Predicted label</p>



[표 9] 최종 분석 모델 테스트 성능 지표

○ 테스트 데이터 예측 결과

배정번호	test	pred	배정번호	test	pred
130675	0	0	127525	1	1
103719	1	1	128388	0	1
114449	0	0	129121	0	0
114453	1	1	130332	1	1
119448	1	1	131938	1	1
120395	1	1	135029	1	1
120867	1	1	135615	1	1
121210	1	1	135704	1	1
123527	1	1	139417	1	1
123708	1	1	140920	0	0
124532	1	1	143370	1	1
124585	1	1	143950	0	0
124960	1	1	144060	1	1
126069	0	0	144308	1	1
126519	1	1	144771	1	0
126569	1	1	146705	1	1
146767	1	1	147546	1	1

[표 10] 최종 분석 모델 테스트 데이터 예측 결과

○ 분석 결과 시사점

본 모델은 전체 테스트 데이터에 대하여 약 92%의 높은 정확도로 분류하였다. '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 판단된다. 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 나타낸다.

□ 중소제조기업에 미치는 파급효과

○ 분석 모델 기대효과

위의 분석 모델은 공정 조건에 따른 품질 예측을 수행하여 문제를 사전에 예방하고 제품 품질을 향상시킬 수 있다. 동시에 불량품 처리 및 재작업 비용을 최소화하여 원가 절감을 실현하며, 저렴한 원가와 높은 품질의 제품을 고객에게 제공하여 고객 만족도를 높일 수 있다.

또한, 데이터 수집, 분석, 품질 모니터링을 자동화함으로써 공정 과정을 효과적으로 제어할 수 있으며, 제품 생산 공정의 효율성을 향상시킨다.

이 모델을 적용할 경우 주기적인 점검에 의존하지 않고 실시간으로 점검이 이루어질 것이다. 이전의 열처리 공정에서는 불량 발생 시 연속적인 불량품 생산과 생산 지연이 발생했지만, 분석 모델을 통해 이러한 문제가 상당 부분 해결될 것으로 기대된다.

○ 분석 모델 확장 가능성

- 타 공정으로의 확장 가능성

위의 분석 모델 알고리즘은 다양한 제조 공정에서 중요한 역할을 할 수 있다.

반도체 에칭 공정은 열을 이용하여 표면에 있는 불필요한 물질을 제거하거나 원하는 패턴을 형성한다. 반도체 제조에서는 미세한 결함이라도 큰 문제로 이어질 수 있는데, 이런 상황에서 분석 모델을 활용하여 불량을 예측하고 품질과 생산성을 향상시킬 수 있다.

금속 가공 및 주조 공정에서는 고온, 고압 등 다양한 환경 조건으로 인해 불량이 발생할 수 있다. 분석 모델을 활용하여 데이터를 분석하고 불량을 사전에 예측하여 품질 수준을 높일 수 있다.

이 외에도 화학 공정, 약물 제조 공정 등 다양한 제조 공정 과정에 위의 분석 모델 알고리즘이 적용될 수 있으며 주요 원인 인자를 추출하여 불량을 예측할 수 있다. 이러한 적용 가능성을 통해 분석 모델은 타 공정으로의 확장 가능하다.

- 타 분야로의 확장 가능성

열처리 기술은 자동차뿐만 아니라 산업 기계, 철도 차량, 금형, 전기, 전자, 건설 중장비 등 다양한 분야에 광범위하게 활용되고 있다. 분석 모델의 알고리즘을 다른 분야의 데이터에 적용함으로써 각 분야에 적합한 모델을 구축할 수 있을 것으로 예상된다.

이러한 분석 모델은 제품 품질 향상, 불량품 예방, 자동화된 생산 프로세스를 제공하며 기업은 경쟁력을 강화하고 고객 만족도를 높일 수 있다.

□ 참고문헌

- [1] G. E. Dieter, “금속강도학“, 회중당, 2012.
- [2] Tianqi Chen, Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System“, arXiv:1603.02754, 2016.
- [3] Sikander, R., Ghulam, A. & Ali, F. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. Sci Rep 12, 5505 (2022). <https://doi.org/10.1038/s41598-022-09484-3>