

# 머신러닝 기법을 활용한 사출공정의 불량 원인 주요 인자 선별 및 품질 예측 모델 개발

2021년 K-인공지능 제조데이터 분석 경진대회

팀명 : M.O

팀원 : 오형택, 구호근, 곽주헌

# CONTENTS

(1) 문제정의

(2) 사출성형 데이터 전처리 및 탐색적 데이터 분석

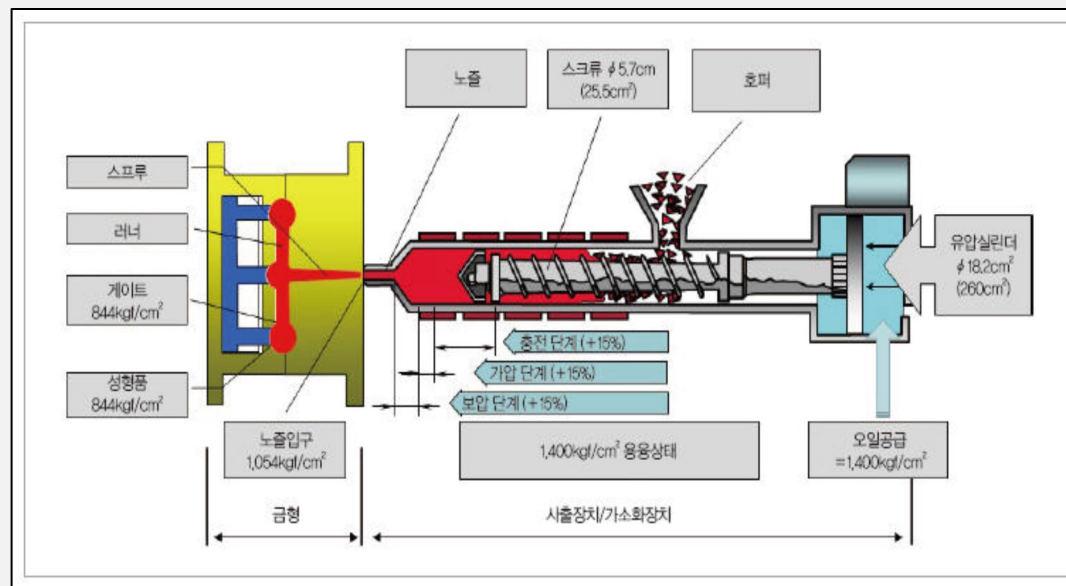
(3) 분석모델 개발

(4) 분석결과 및 시사점

# 문제정의

## ■ 사출성형 정의

- 플라스틱(열경화성 수지, 열가소성 수지)을 원하는 형태의 성형품으로 만드는 기술
- 성형의 사이클이 짧고 능률이 좋아 대량생산이 가능하며 복잡한 모양까지 생산 가능
- 사출성형은 사출장치/가소화장치, 금형 등 다양한 단계를 통과하여 공정이 수행됨



< 사출성형 공정 >

# 문제정의

## ■ 사출성형 공정의 문제점

- 공정내에 간혹 성형 불량품이 발생하지만 주요 원인에 대해 알 수 없음
- 불량품 발생시 불량 원인에 대한 작업자의 빠른 조취 필요
- 사출품의 양품/불량품 검사는 작업자의 개입이 필요

## ■ 사출성형 공정의 개선 방안

- 머신러닝 모델 중 트리기반 모델과 부스팅 기반 모델을 사용하여 불량품을 예측하는 주요 인자 파악
- 사출품의 양품/불량품을 구별해내는 인공지능 모델 구축을 통한 공정의 효율적 운영 및 품질 검사 비용 절감

# 사출성형 데이터 전처리

## ■ 사출성형기 데이터셋 전처리 과정

- 사출품에 대한 품질 불량 여부를 알 수 있는 "경진대회용 사출성형기 데이터셋\_labeled.csv" 활용
- 사출공정의 45개의 변수와 7996개의 데이터셋으로 구성
- 생산한 사출기 모델명(EQUIP\_NAME) 변수에서 "650톤-우진2호기" 데이터만 선별
- 제품의 이름에 따라 CN7&RG3 데이터셋, CN7 데이터셋, RG3 데이터셋으로 구분
- 시간변수, 0으로만 채워진 변수 등 20개의 컬럼 제거  
['\_id', 'TimeStamp', 'PART\_FACT\_PLAN\_DATE', 'Reason', 'PART\_FACT\_SERIAL', 'PART\_NAME', 'EQUIP\_CD', 'EQUIP\_NAME', 'Barrel\_Temperature\_7', 'Switch\_Over\_Position', 'Mold\_Temperature\_1', 'Mold\_Temperature\_2', 'Mold\_Temperature\_5', 'Mold\_Temperature\_6', 'Mold\_Temperature\_7', 'Mold\_Temperature\_8', 'Mold\_Temperature\_9', 'Mold\_Temperature\_10', 'Mold\_Temperature\_11', 'Mold\_Temperature\_12']
- 1개의 종속변수(PassOrFail)와 24개의 독립변수 사용

```
# 생산한 사출기 모델명 갯수 비교
label[ 'EQUIP_NAME' ].value_counts()

650톤-우진2호기      7992
1800TON-우진         2
650톤-우진           2
Name: EQUIP_NAME, dtype: int64
```

< 생산한 사출기 모델명 갯수 비교 >

# 사출성형 데이터 전처리

## ■ 사출성형기 데이터셋 전처리 과정

Data columns (total 25 columns):			
#	Column	Non-Null Count	Dtype
0	PassOrFail	7992 non-null	object
1	Injection_Time	7992 non-null	float64
2	Filling_Time	7992 non-null	float64
3	Plasticizing_Time	7992 non-null	float64
4	Cycle_Time	7992 non-null	float64
5	Clamp_Close_Time	7992 non-null	float64
6	Cushion_Position	7992 non-null	float64
7	Plasticizing_Position	7992 non-null	float64
8	Clamp_Open_Position	7992 non-null	float64
9	Max_Injection_Speed	7992 non-null	float64
10	Max_Screw_RPM	7992 non-null	float64
11	Average_Screw_RPM	7992 non-null	float64
12	Max_Injection_Pressure	7992 non-null	float64
13	Max_Switch_Over_Pressure	7992 non-null	float64
14	Max_Back_Pressure	7992 non-null	float64
15	Average_Back_Pressure	7992 non-null	float64
16	Barrel_Temperature_1	7992 non-null	float64
17	Barrel_Temperature_2	7992 non-null	float64
18	Barrel_Temperature_3	7992 non-null	float64
19	Barrel_Temperature_4	7992 non-null	float64
20	Barrel_Temperature_5	7992 non-null	float64
21	Barrel_Temperature_6	7992 non-null	float64
22	Hopper_Temperature	7992 non-null	float64
23	Mold_Temperature_3	7992 non-null	float64
24	Mold_Temperature_4	7992 non-null	float64

< 사용 데이터 셋 정보 >

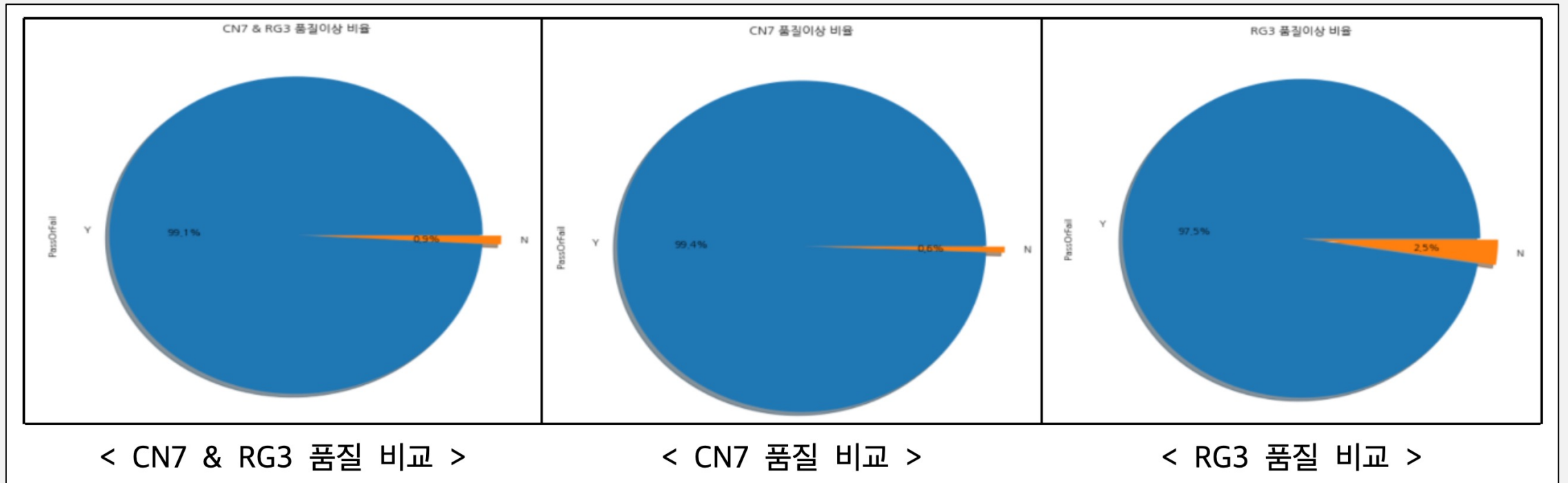
## ■ 사용 데이터 셋 모양 확인

- CN7&RG3 데이터 셋 : (7992, 25)
- CN7 데이터 셋 : (6736, 25)
- RG3 데이터 셋 : (1256, 25)

# 탐색적 데이터 분석

## ■ 사출성형기 데이터셋 탐색적 데이터 분석

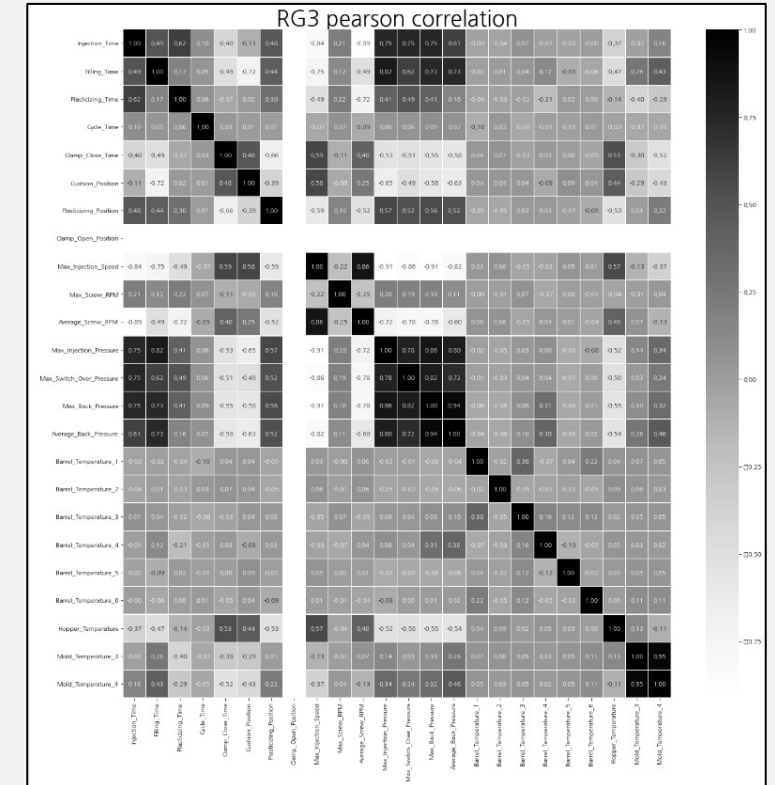
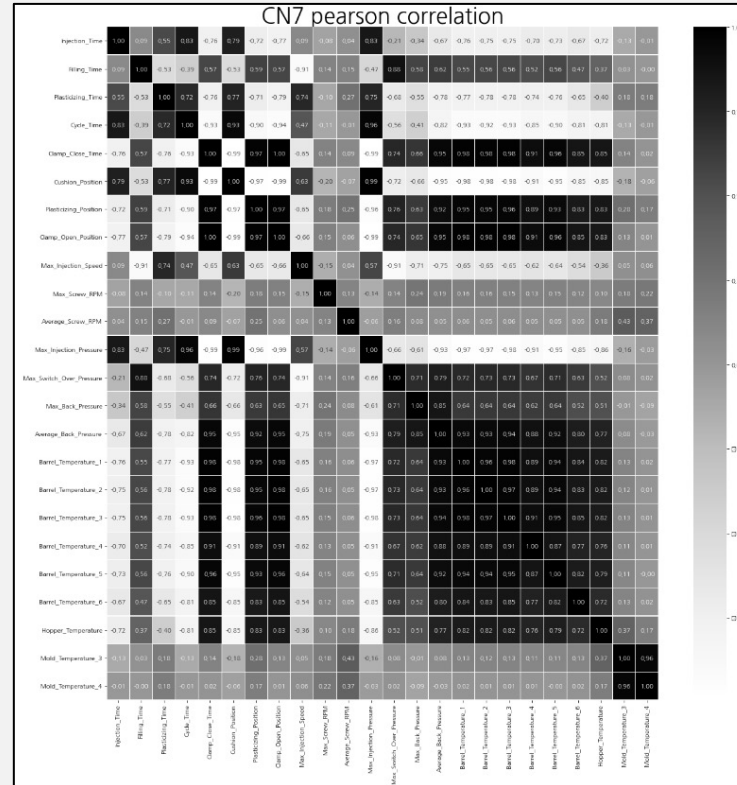
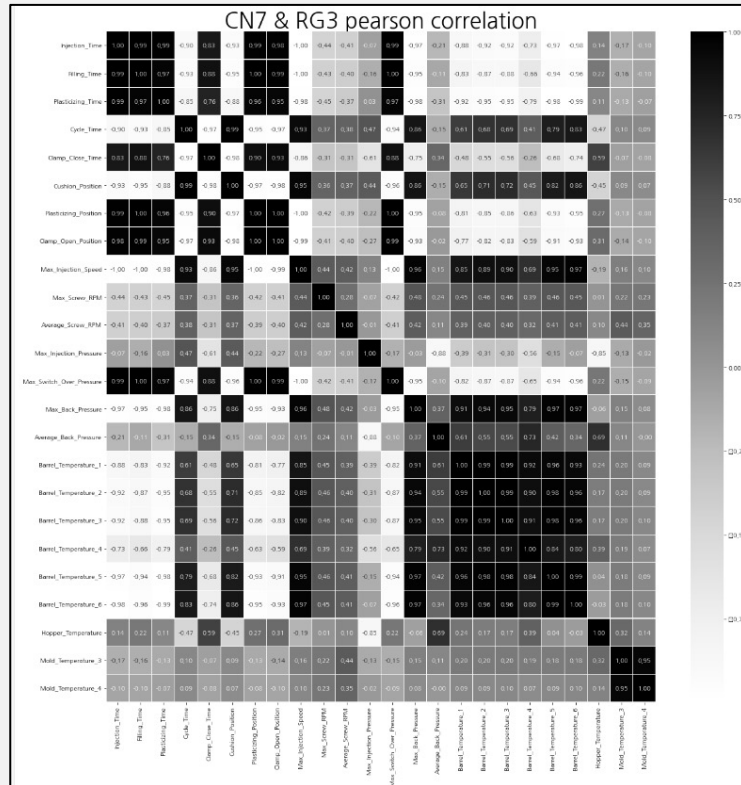
- 사출품 양품과 불량품 비율 확인
- 양품과 불량품에 대한 심각한 클래스 불균형 확인



# 탐색적 데이터 분석

## ■ 사출성형기 데이터셋 탐색적 데이터 분석

- 각 데이터별 피어슨 유사도 확인
- CN7&RG3, CN7 데이터 셋 "Barrel\_Temperature" 변수간 높은 상관성
- RG3 데이터 셋 "Pressure" 변수간 높은 상관성

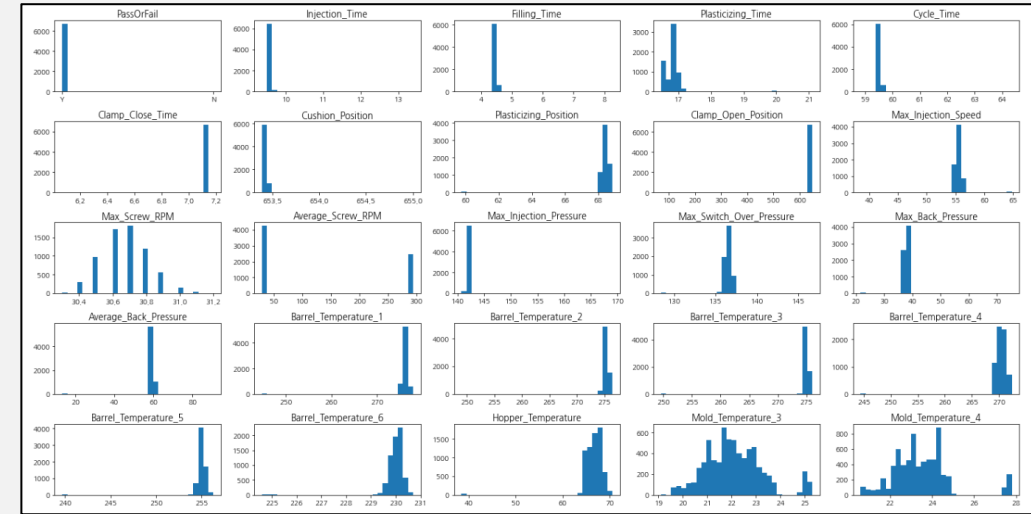




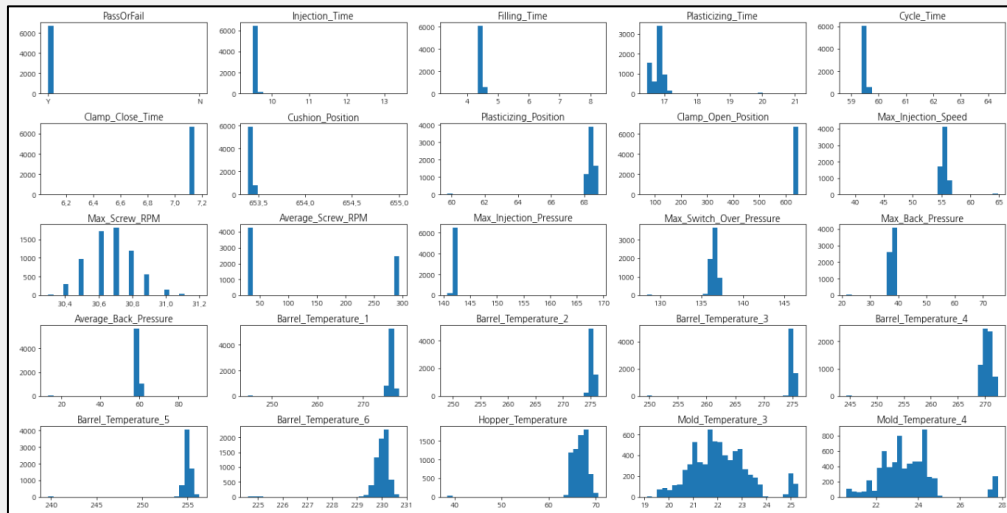
# 탐색적 데이터 분석

## ■ 사출성형기 데이터셋 탐색적 데이터 분석

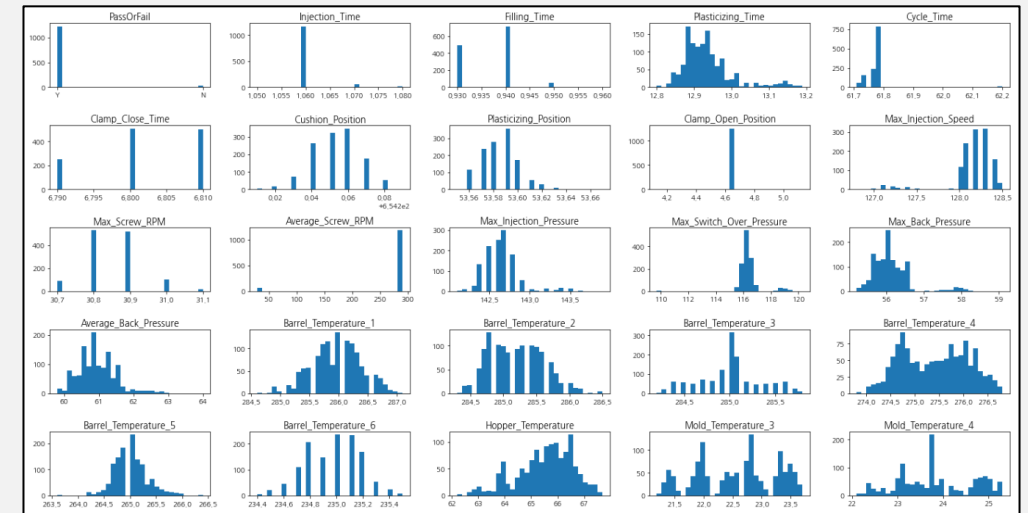
- 각 데이터별 주요 인자 분포 확인
- CN7, RG3 데이터 셋 뚜렷한 분포차이 존재하지 않음



< CN7&RG3 주요 인자 분포 >



< CN7 주요 인자 분포 >

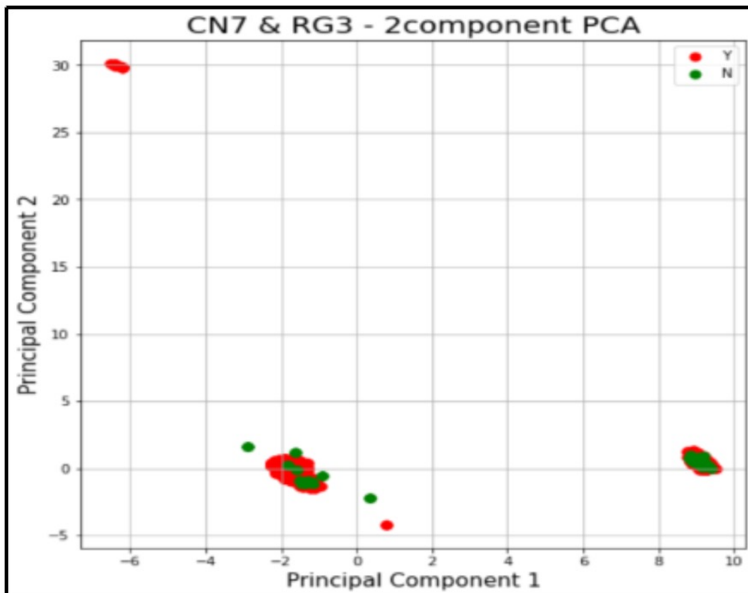


< RG3 주요 인자 분포 >

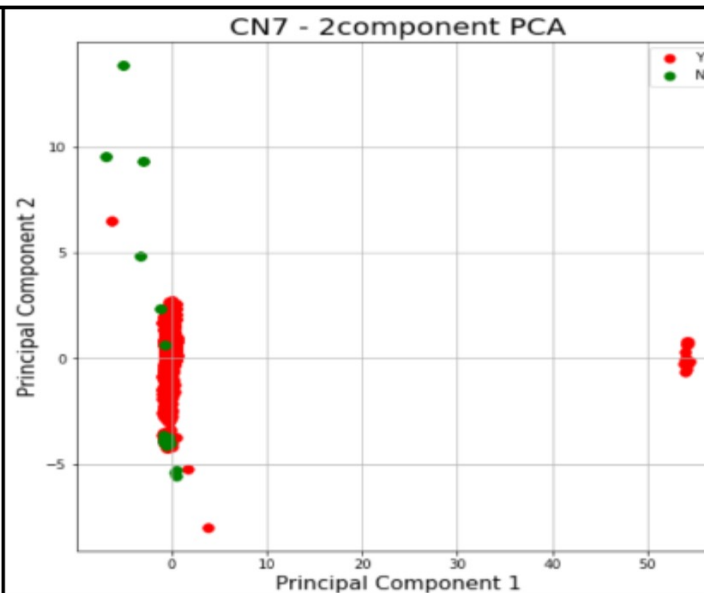
# 탐색적 데이터 분석

## ■ 사출성형기 데이터셋 탐색적 데이터 분석

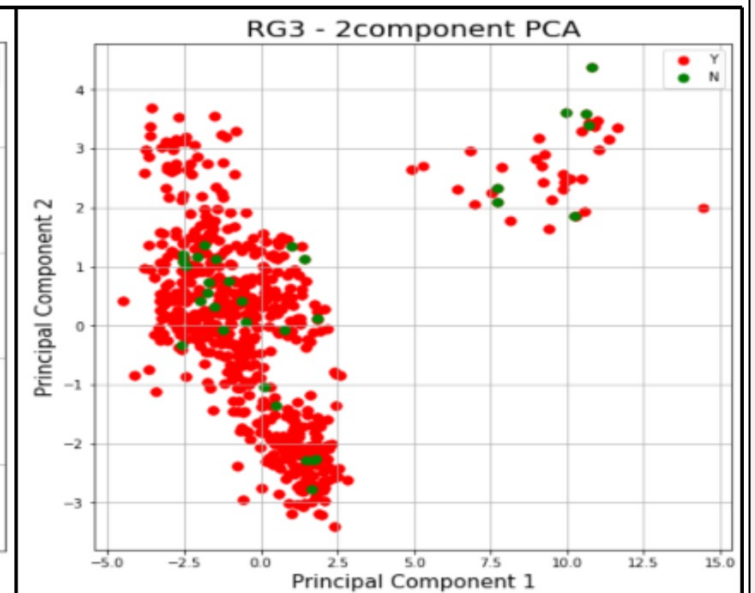
- PCA를 통한 주요 인자 분포 확인
- PCA를 통한 양품과 불량품의 분포 차이를 뚜렷하게 구분할 수 없음



< CN7 & RG3 PCA 분포 >



< CN7 PCA 분포 >



< RG3 PCA 분포 >

# 사출성형 데이터 분석모델 개발

## ■ 사출공정의 불량 원인 인자 파악을 위한 머신러닝 모델

- 의사결정나무 기반의 앙상블 모델의 경우 지니 불순도(Gini impurity)를 통해 변수의 중요도를 알 수 있음
- 부스팅 기반의 모델 역시 노드가 특정 변수로 분기되었을 때 얻는 성능 상의 이득(gain)을 통해 변수의 중요도를 계산할 수 있다
- 클래스 불균형이 존재하는 사출성형 데이터셋에서 변수의 중요도를 계산할 경우 클래스가 많은 양품 데이터에 치우쳐 불량 원인에 대한 주요 요인을 판단 불가
- 양품데이터에서 불량품의 수 만큼 랜덤으로 과소표본추출을 진행하여, 양품과 불량품의 비율을 동일하게 맞추어줌
- Scikit-learn 패키지를 활용하여 앙상블기반 모델과 부스팅 기반 모델 구축 후 변수의 중요도 확인

# 사출성형 데이터 분석결과

## ■ 사출공정의 불량 원인에 대한 주요 인자 파악

- CN7 & RG3 제품 사출품의 양품 데이터를 과소표본추출하여 양품과 불량품을 동일한 비율로 맞춤
- 의사결정나무 > 랜덤포레스트 > XGBoost > LightGBM 순서로 모델 성능 차이를 보임

모델	정확도	정밀도	재현율	F1	ROC-AUC
랜덤포레스트	0.860	0.778	<b>1.0</b>	0.875	0.864
의사결정 나무	<b>0.884</b>	<b>0.833</b>	<b>0.952</b>	<b>0.889</b>	<b>0.885</b>
XGBoost	0.860	0.826	0.905	0.864	0.861
LightGBM	0.837	0.792	0.905	0.844	0.839

< CN7 & RG3 불량 원인 인자 파악을 위한 모델 비교 >

# 사출성형 데이터 분석결과

## ■ 사출공정의 불량 원인에 대한 주요 인자 파악

- 변수 중요도를 활용한 불량품을 결정하는 주요 인자 확인
- Average\_Back\_Pressure과 Plasticizing\_Position이 불량품을 예측하는 주요 요인
- 해당 주요 인자 양품과 불량품에 대한 평균, 표준편차 등의 수치적 차이 나타남

< 표 - CN7 & RG3 불량원인 주요인자 TOP-3 >

	TOP-1	TOP-2	TOP-3
랜덤포레스트	Average_Back_Pressure	Plasticizing_Position	Barrel_Temperature_5
의사결정나무	Average_Back_Pressure	Plasticizing_Position	Mold_Temperature_3
XGBoost	Average_Back_Pressure	Plasticizing_Position	Max_Injection_Pressure
LightGBM	Plasticizing_Position	Plasticizing_Time	Mold_Temperature_4

< 표 - CN7 & RG3 제품 불량원인 주요인자 기술통계량 >

	평균	표준편차	최솟값	최댓값	25%	50%	75%
양품 Average_Back_Pressure	59.57	3.25	13.3	90.8	59.4	59.6	59.9
불량품 Average_Back_Pressure	62.42	6.38	58.7	87.1	59.7	60.6	62.0
양품 Plasticizing_Position	66.09	5.38	53.55	68.86	68.25	68.34	68.49
불량품 Plasticizing_Position	61.88	7.57	53.56	68.86	53.59	68.58	68.61

# 사출성형 데이터 분석결과

## ■ 사출공정의 불량 원인에 대한 주요 인자 파악

- CN7 제품 사출품의 양품 데이터를 과소표본추출하여 양품과 불량품을 동일한 비율로 맞춤
- 랜덤포레스트 = XGBoost > 의사결정나무 = LightGBM 순서로 모델 성능 차이를 보임

모델	정확도	정밀도	재현율	F1	ROC-AUC
랜덤포레스트	<b>0.917</b>	<b>0.857</b>	<b>1.0</b>	<b>0.923</b>	<b>0.917</b>
의사결정 나무	0.875	0.8	<b>1.0</b>	0.889	0.875
XGBoost	<b>0.917</b>	<b>0.857</b>	<b>1.0</b>	<b>0.923</b>	<b>0.916</b>
LightGBM	0.875	0.8	<b>1.0</b>	0.889	0.875

< CN7 불량 원인 인자 파악을 위한 모델 비교 >

# 사출성형 데이터 분석결과

## ■ 사출공정의 불량 원인에 대한 주요 인자 파악

- 변수 중요도를 활용한 불량품을 결정하는 주요 인자 확인
- Plasticizing\_Position 가장 중요한 변수로 선택하였고, 그외 Max\_Injection\_Speed, Mold\_Temperature 등
- Plasticizing\_Position 양품과 불량품 수치적 차이 없음
- 나머지 주요 인자인 Max\_Injection\_Speed 양품과 불량품에 대한 평균, 표준편차 등의 수치적 차이 나타남

< 표 - CN7 불량원인 주요인자 TOP-3 >

	TOP-1	TOP-2	TOP-3
랜덤포레스트	Plasticizing_Position	Mold_Temperature_4	Max_Injection_Speed
의사결정나무	Plasticizing_Position	Cycle_Time	Barrel_Temperature_1
XGBoost	Plasticizing_Position	Average_Back_Pressure	Max_Injection_Speed
LightGBM	Plasticizing_Position	Barrel_Temperature_5	Hopper_Temperature_5

< 표 - CN7 제품 불량원인 주요인자 기술통계량 >

	평균	표준편차	최솟값	최댓값	25%	50%	75%
양품							
Plasticizing_Position	68.38	0.65	59.76	68.86	68.32	68.36	68.51
불량품							
Plasticizing_Position	68.69	0.15	68.36	68.86	68.59	68.61	68.84
양품							
Max_Injection_Speed	55.55	0.82	49.3	64.8	55.1	55.4	55.9
불량품							
Max_Injection_Speed	51.6	6.77	38.5	45.4	53.3	55.5	60.9

# 사출성형 데이터 분석결과

## ■ 사출공정의 불량 원인에 대한 주요 인자 파악

- RG3 제품 사출품의 양품 데이터를 과소표본추출하여 양품과 불량품을 동일한 비율로 맞춤
- RG3 데이터 셋에서는 머신러닝 모델이 양품과 불량품을 구분해내지 못함
- 사출품의 불량을 야기하는 주요인자 확인 불가

< RG3 불량 원인 인자 파악을 위한 모델 비교 >					
모델	정확도	정밀도	재현율	F1	ROC-AUC
랜덤포레스트	0.4	0.375	<b>0.3</b>	0.333	0.404
의사결정 나무	0.3	0.25	0.2	0.222	0.3
XGBoost	0.3	0.25	0.2	0.22	0.303
lightGBM	<b>0.5</b>	<b>0.5</b>	<b>0.3</b>	<b>0.375</b>	<b>0.5</b>



# 사출성형 데이터 분석모델 개발

## ■ 사출품 품질 예측을 위한 인공지능 모델 개발

- 사출품 품질을 예측하기 위해 8개의 예측모델 구축
- 의사결정나무, 랜덤포레스트, 서포트 벡터 머신, XGBoost, LightGBM, MLP(1-Layer), Deep-MLP(3-Layer), 오토인코더 사용
- 위 모델 중 성능이 우수한 모델에 대하여 SMOTE를 적용한 데이터 오버 샘플링 기법 적용
- 학습 셋 & 테스트 셋을 7:3으로 분리
- StandardScaler를 통한 데이터 표준화 진행
- Accuracy, Recall, Precision, F1-score, ROC-AUC 평가 지표 활용

# 사출성형 데이터 분석결과

## ■ 사출품 품질 예측을 위한 인공지능 모델 개발

- CN7&RG3 제품 인공지능 모델 구축 후 평가
- 의사결정나무가 가장 좋은 성능을 나타냄
- SMOTE를 적용하여 불량품 데이터를 오버샘플링 후 모델 학습 결과 성능이 현저히 떨어짐

< CN7 & RG3 사출품 품질 예측을 위한 인공지능 모델 >					
모델	정확도	정밀도	재현율	F1	ROC-AUC
랜덤포레스트	0.988	0.316	<b>0.286</b>	0.3	0.640
의사결정 나무	<b>0.994</b>	<b>1.0</b>	<b>0.286</b>	<b>0.445</b>	<b>0.643</b>
서포트 벡터 머신	0.993	<b>1.0</b>	0.190	0.320	0.595
XGBoost	0.988	0.316	<b>0.286</b>	0.3	0.640
LightGBM	0.988	0.316	<b>0.286</b>	0.3	0.640
MLP	0.993	<b>1.0</b>	0.238	0.385	0.620
Deep-MLP	0.992	0.6	<b>0.286</b>	0.387	0.642
오토인코더	0.819	0.0	0.0	0.0	0.421
의사결정 나무 (SMOTE 적용)	0.009	0.009	1.0	0.017	0.5

# 사출성형 데이터 분석결과

## ■ 사출품 품질 예측을 위한 인공지능 모델 개발

- CN7 제품에서 인공지능 모델 구축 후 평가
- 의사결정나무가 가장 좋은 성능을 나타내며, CN7&RG3 환경보다 좋은 성능 개선이 나타남
- SMOTE를 적용하여 불량품 데이터를 오버샘플링 후 모델 학습 결과 성능이 현저히 떨어짐

< CN7 사출품 품질 예측을 위한 인공지능 모델 개발 >					
모델	정확도	정밀도	재현율	F1	ROC-AUC
랜덤포레스트	0.996	0.643	<b>0.75</b>	0.692	0.874
의사결정 나무	<b>0.999</b>	<b>1.0</b>	<b>0.75</b>	<b>0.857</b>	<b>0.875</b>
서포트 벡터 머신	0.998	<b>1.0</b>	0.583	0.737	0.792
XGBoost	0.996	0.643	<b>0.75</b>	0.682	0.874
LightGBM	0.996	0.643	<b>0.75</b>	0.692	0.874
MLP	0.998	0.818	<b>0.75</b>	0.783	<b>0.875</b>
Deep-MLP	0.996	0.6	<b>0.75</b>	0.667	0.874
오토인코더	0.976	0.211	0.103	0.138	0.548
의사결정 나무 (SMOTE 적용)	0.006	0.006	1.0	0.012	0.5

# 사출성형 데이터 분석결과

## ■ 사출품 품질 예측을 위한 인공지능 모델 개발

- RG3 제품에서 인공지능 모델 구축 후 평가
- 인공지능 모델이 사출품의 품질을 예측하지 못함

< RG3 사출품 품질 예측을 위한 인공지능 모델 개발 >					
모델	정확도	정밀도	재현율	F1	ROC-AUC
랜덤포레스트	0.966	0.0	0.0	0.0	0.496
의사결정 나무	<b>0.973</b>	0.0	0.0	0.0	<b>0.5</b>
서포트 벡터 머신	0.973	0.0	0.0	0.0	<b>0.5</b>
XGBoost	0.966	0.0	0.0	0.0	0.496
LightGBM	0.966	0.0	0.0	0.0	0.496
MLP	0.971	0.0	0.0	0.0	0.499
Deep-MLP	0.966	0.0	0.0	0.0	0.496
오토인코더	0.873	<b>0.047</b>	<b>0.031</b>	<b>0.037</b>	0.488
의사결정 나무 (SMOTE 적용)	0.973	0.0	0.0	0.0	0.5

## 결론 및 시사점

- 평균 압력과 계량완료위치 등이 사출품 불량 원인을 야기하는 주요 요인으로 나타남
- 의사결정나무 모델은 사출공정의 품질을 예측하는데 우수한 성능을 나타냄
- 다양한 제조 공정의 불명확한 불량 원인을 알아보기 위해 변수 중요도 파악 방법은 품질의 저하를 야기하는 주요 인자 관리할 수 있음
- 의사결정나무 모델을 사용하여 다양한 제조 공정의 품질 이상 문제를 효율적으로 관리 가능

감사합니다