CMM 데이터 이상치 탐지 딥러닝 모듈 개발 [7주차]

AICMM팀 (김지선, 김예령, 백수민)

발표일자: 2024-04-08

실증적AI프로젝트 금주 활동계획 (7주차)

주제: CMM 데이터의 이상치 탐지 딥러닝 모듈 개발

<table>
<tr><td>금주
활동계획</td><td colspan="3">1. CMM 데이터에 더 다양한 ML 모델 실험
2. GNN(Graph Neural Network) 논문 리딩 및 정리</td></tr>
<tr><td></td><td>팀장 (김지선)</td><td>팀원1 (김예령)</td><td>팀원2 (백수민)</td></tr>
<tr><td>금주
개인별
활동계획</td><td>1. CMM 데이터 ML에 적용
• CMM 측정 구성요소와 동작원리에 대해 공부
• ML에 적용</td><td>2. CMM 데이터 전처리
• 데이터를 머신러닝 모델에 입력할 수 있는 형태로 데이터 전처리
• CMM 데이터 특성 정리.</td><td>3. GNN 논문 리딩 및 정리
• GNN 관련 논문 수집
• GNN 관련 논문 리딩 및 정리</td></tr>
<tr><td>차주
활동계획</td><td colspan="3">1. GNN(Graph Neural Network) 이상치처리 모델 공부
2. GNN(Graph Neural Network) 이상치처리 모델 구현 실습</td></tr>
</table>

7주차 진행사항

1) CMM(Coordinate Measuring Machine) 데이터 전처리

2) 머신러닝을 활용한 CMM 데이터 이상치처리

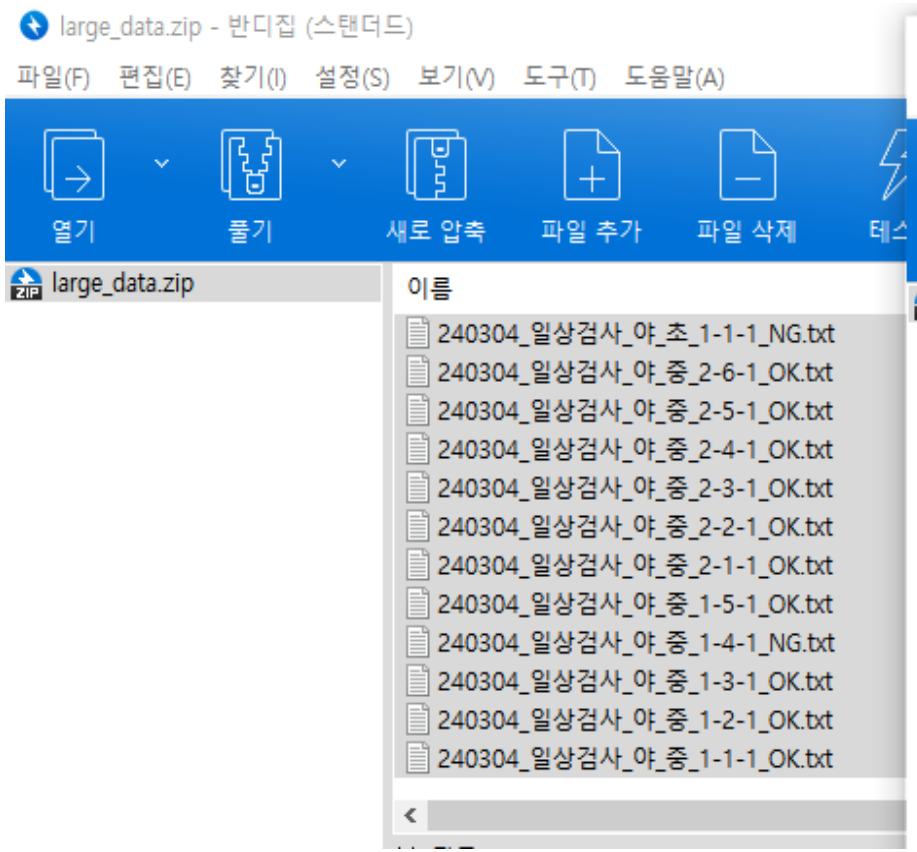
- 랜덤 포레스트 (Random forest)
- 로지스틱 회귀 (Logistic regression)

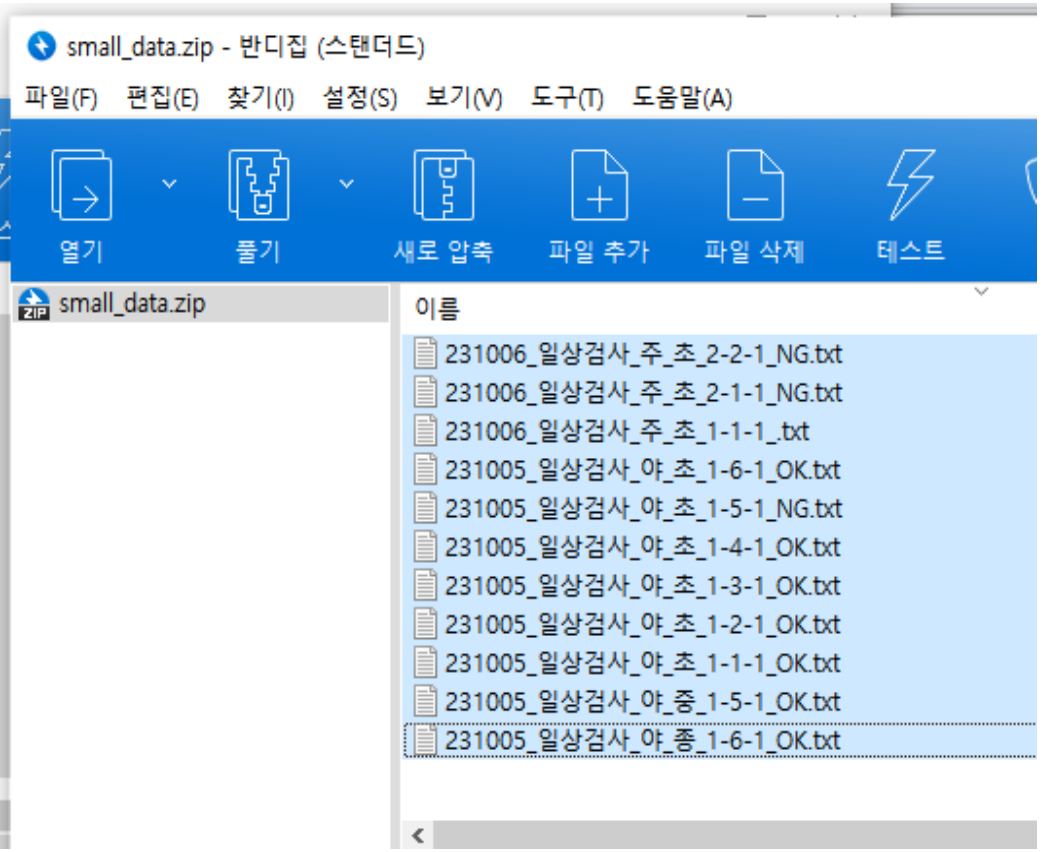
데이터 전처리 워크플로우

- 1단계: 데이터 구조 파악 및 통합
- 2단계: 텍스트 파일을 엑셀 형태로 변환하기 위해 파싱(Parsing)
- 3단계: 모든 파일 엑셀로 재구성

데이터 구조 파악

- 회사로부터 받은 데이터셋은 zip 파일 형태의 두 가지 폴더로 구성
  - 폴더: small_data.zip, large_data.zip)





샘플 데이터셋

- 데이터 통합을 위해 이를 한 폴더(Sample_data)로 통합

구분	형태 (행, 열)	데이터 개수
Small data	(74, 17)	11개
Large data	(76, 17)	137개

구분	형태 (행, 열)	데이터 개수
Sample_data	(76, 17)	148개

데이터 압축 해제 중 해결된 오류

- txt를 csv 파일로 변환 완료! 하지만 한글 깨짐
- 압축을 푸는 과정에서 한글 깨짐 확인 -> 한국어로 파일명 압축 푼 후 csv로 변환

이름

231005_[illegible]_1-1-1_OK.csv
231005_[illegible]_1-1-1_OK.txt
231005_[illegible]_1-2-1_OK.csv
231005_[illegible]_1-2-1_OK.txt
231005_[illegible]_1-3-1_OK.csv
231005_[illegible]_1-3-1_OK.txt
231005_[illegible]_1-4-1_OK.csv
231005_[illegible]_1-4-1_OK.txt

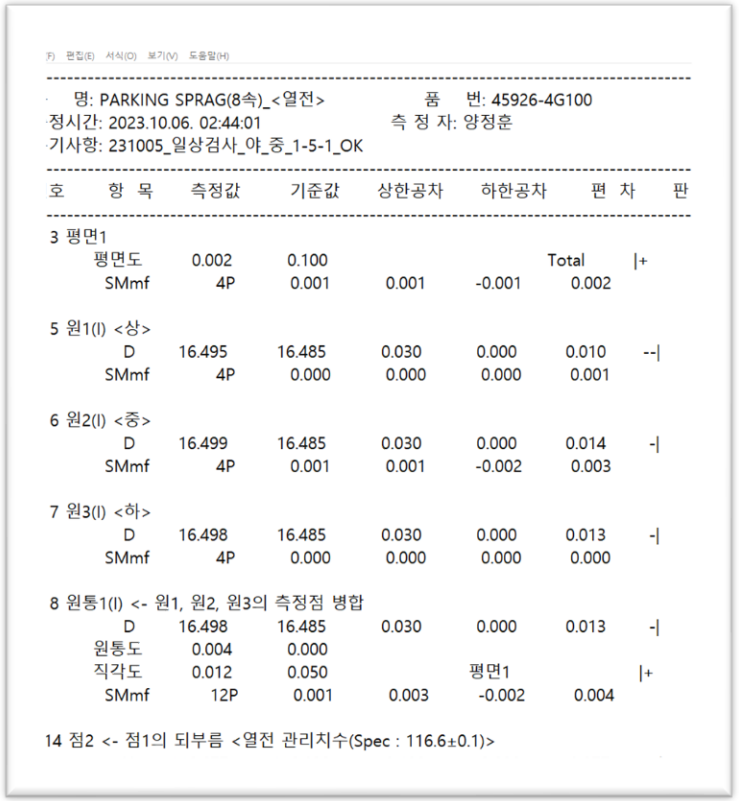
<오류 해결 전>

231005_일상검사_야_초_1-1-1_OK.csv
231005_일상검사_야_초_1-1-1_OK.txt
231005_일상검사_야_초_1-2-1_OK.csv
231005_일상검사_야_초_1-2-1_OK.txt
231005_일상검사_야_초_1-3-1_OK.csv
231005_일상검사_야_초_1-3-1_OK.txt
231005_일상검사_야_초_1-4-1_OK.csv
231005_일상검사_야_초_1-4-1_OK.txt

<오류 해결 후>

파싱 함수 구현

- 데이터를 머신러닝 모델의 입력으로 사용하기 위해 텍스트(.txt)를 엑셀(.csv) 형태로 변환 필요

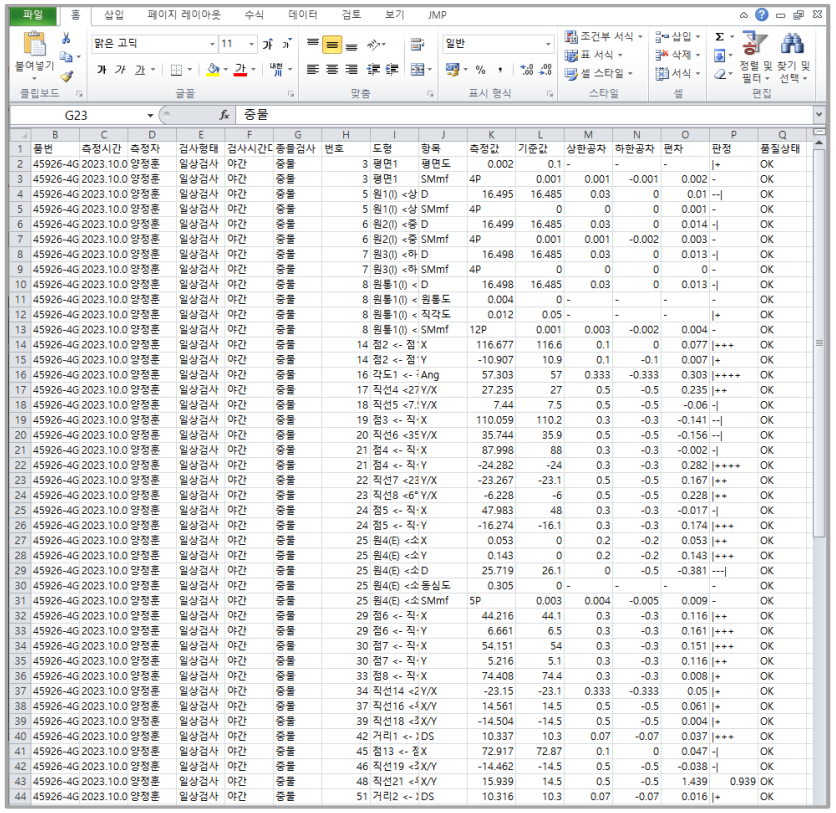


F) 편집(E) 서식(O) 보기(V) 도움말(H)

명: PARKING SPRAG(8속)_<열전> 품 번: 45926-4G100
정시간: 2023.10.06. 02:44:01 측 정 자: 양정훈
기사항: 231005_일상검사_야_중_1-5-1_OK

호	항 목	측정값	기준값	상한공차	하한공차	편 차	판
3	평면1						
	평면도	0.002	0.100			Total	|+
	SMmf	4P	0.001	0.001	-0.001	0.002	
5	원1(I) <상>						
	D	16.495	16.485	0.030	0.000	0.010	--|
	SMmf	4P	0.000	0.000	0.000	0.001	
6	원2(I) <중>						
	D	16.499	16.485	0.030	0.000	0.014	-|
	SMmf	4P	0.001	0.001	-0.002	0.003	
7	원3(I) <하>						
	D	16.498	16.485	0.030	0.000	0.013	-|
	SMmf	4P	0.000	0.000	0.000	0.000	
8	원통1(I) <- 원1, 원2, 원3의 측정점 병합						
	D	16.498	16.485	0.030	0.000	0.013	-|
	원통도	0.004	0.000				
	직각도	0.012	0.050		평면1		|+
	SMmf	12P	0.001	0.003	-0.002	0.004	
14	점2 <- 점1의 되부름 <열전 관리치수(Spec : 116.6±0.1)>						

<텍스트 데이터>



품번	측정시간	측정자	검사형태	검사시간	종물검사	번호	도형	항목	측정값	기준값	상한공차	하한공차	편차	판정	품질상태
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	3	평면1	평면도	0.002	0.1	-	-	-	|+	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	3	평면1	SMmf	4P	0.001	0.001	-0.001	0.002	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	5	원1(I) <상	D	16.495	16.485	0.03	0	0.01	--|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	5	원1(I) <상	SMmf	4P	0	0	0	0.001	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	6	원2(I) <중	D	16.499	16.485	0.03	0	0.014	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	6	원2(I) <중	SMmf	4P	0.001	0.001	-0.002	0.003	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	7	원3(I) <하	D	16.498	16.485	0.03	0	0.013	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	7	원3(I) <하	SMmf	4P	0	0	0	0	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	8	원통1(I) <	D	16.498	16.485	0.03	0	0.013	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	8	원통1(I) <	원통도	0.004	0	-	-	-	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	8	원통1(I) <	직각도	0.012	0.05	-	-	-	|+	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	8	원통1(I) <	SMmf	12P	0.001	0.003	-0.002	0.004	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	14	점2 <- 점	X	116.677	116.6	0.1	0	0.077	|+++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	14	점2 <- 점	Y	-10.907	10.9	0.1	-0.1	0.007	|+	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	16	각도1 <-	Ang	57.303	57	0.333	-0.333	0.303	|++++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	17	직선4 <27	Y/X	27.235	27	0.5	-0.5	0.235	|++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	18	직선5 <7	Y/X	7.44	7.5	0.5	-0.5	-0.06	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	19	점3 <- 직	X	110.059	110.2	0.3	-0.3	-0.141	--|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	20	직선6 <35	Y/X	35.744	35.9	0.5	-0.5	-0.156	--|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	21	점4 <- 직	X	87.998	88	0.3	-0.3	-0.002	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	21	점4 <- 직	Y	-24.282	-24	0.3	-0.3	0.282	|++++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	22	직선7 <23	Y/X	-23.267	-23.1	0.5	-0.5	0.167	|++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	23	직선8 <6°	Y/X	-6.228	-6	0.5	-0.5	0.228	|++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	24	점5 <- 직	X	47.983	48	0.3	-0.3	-0.017	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	24	점5 <- 직	Y	-16.274	-16.1	0.3	-0.3	0.174	|+++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	25	원4(E) <소	X	0.053	0	0.2	-0.2	0.053	|++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	25	원4(E) <소	Y	0.143	0	0.2	-0.2	0.143	|+++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	25	원4(E) <소	D	25.719	26.1	0	-0.5	-0.381	---|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	25	원4(E) <소	동심도	0.305	0	-	-	-	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	25	원4(E) <소	SMmf	5P	0.003	0.004	-0.005	0.009	-	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	29	점6 <- 직	X	44.216	44.1	0.3	-0.3	0.116	|++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	29	점6 <- 직	Y	6.661	6.5	0.3	-0.3	0.161	|+++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	30	점7 <- 직	X	54.151	54	0.3	-0.3	0.151	|+++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	30	점7 <- 직	Y	5.216	5.1	0.3	-0.3	0.116	|++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	33	점8 <- 직	X	74.408	74.4	0.3	-0.3	0.008	|+	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	34	직선14 <2	Y/X	-23.15	-23.1	0.333	-0.333	0.05	|+	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	37	직선16 <4	X/Y	14.561	14.5	0.5	-0.5	0.061	|+	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	39	직선18 <3	X/Y	-14.504	-14.5	0.5	-0.5	0.004	|+	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	42	거리1 <-	DS	10.337	10.3	0.07	-0.07	0.037	|+++	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	45	점13 <- 점	X	72.917	72.87	0.1	0	0.047	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	46	직선19 <3	X/Y	-14.462	-14.5	0.5	-0.5	-0.038	-|	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	48	직선21 <4	X/Y	15.939	14.5	0.5	-0.5	1.439	0.939	OK
45926-4G	2023.10.0	양정훈	일상검사	야간	중물	51	거리2 <-	DS	10.316	10.3	0.07	-0.07	0.016	|+	OK

<엑셀 데이터>

파싱 함수 동작 과정

- 파일에서 데이터를 추출하는 함수(extract_datafram_from_file) 정의
- 텍스트 파일에서 제품명,제품번호,측정항목 등 헤더 정보를 추출
- 해당 데이터 행을 추출하여 행과 열로 구조화 하여 데이터 프레임을 형성

품명

헤더

측정항목

```
품      명: PARKING SPRAG(8속)_<열전>          품      번: 45926-4G100
측정시간: 2023.10.06. 02:44:01                 측 정 자: 양정훈
특기사항: 231005_일상검사_야_중_1-5-1_OK
-----------------------------------------------------------------------------------
번호 도형 항  목      측정값      기준값      상한공차      하한공차      편  차      판  정
-----------------------------------------------------------------------------------
 3 평면1
      평면도      0.002       0.100                                Total      |+
        SMmf        4P        0.001       0.001      -0.001       0.002

 5 원1(I) <상>
          D       16.495      16.485       0.030       0.000       0.010      --|
        SMmf        4P        0.000       0.000       0.000       0.001

 6 원2(I) <중>
          D       16.499      16.485       0.030       0.000       0.014      -|
        SMmf        4P        0.001       0.001      -0.002       0.003

 7 원3(I) <하>
          D       16.498      16.485       0.030       0.000       0.013      -|
        SMmf        4P        0.000       0.000       0.000       0.000
```

텍스트 데이터 전처리 방법

- 데이터 파싱 후 아래 값들을 기준으로 데이터 프레임 형태로 재구성
- 데이터프레임을 엑셀로 저장해서 데이터 변환

구분	명칭
품명	PARKING SPRAG (8속)_<열전>
품번	45926-4G10
측정시간	2023.xx.xx.xx:xx:xx
측정자	OOO
검사형태	일상검사 / 치수보정
검사 시간대	주간/야간
종물검사	초 / 중 / 종물

<기준 정보>

번호	도형	항목	측정값	기준값	·····	판정	품질상태
1	평면1	평면도	0.003	0.1	·····	|+	OK
2	평면1	SMmf	4P	0.001		-	OK
3	원1(I) <상>	D	16.485	16.485		----|	OK
4	원1(I) <상>	SMmf	4P	0.003		-	OK
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

<데이터프레임 형태>

머신러닝을 활용한 CMM 데이터 이상치처리

- 각 CSV 파일에서 도형, 항목에 따른 편차 및 품질 상태 데이터를 가져옴. (cmm_data.csv)
- 편차의 값이 비어져 있으면 각 CSV 파일의 편차 평균 값을 사용함.

이름	수정한 날짜
240304_일상검사_야_중_1-1-1_OK	2024-04-1
240304_일상검사_야_중_1-2-1_OK	2024-04-1
240304_일상검사_야_중_1-3-1_OK	2024-04-1
240304_일상검사_야_중_1-4-1_NG	2024-04-1
240304_일상검사_야_중_1-5-1_OK	2024-04-1
240304_일상검사_야_중_2-1-1_OK	2024-04-1
240304_일상검사_야_중_2-2-1_OK	2024-04-1
240304_일상검사_야_중_2-3-1_OK	2024-04-1
240304_일상검사_야_중_2-4-1_OK	2024-04-1

평면1, 평면도	원1(I) <상>, D	원2(I) <중>, D	원3(I) <하>, D
0.045166667	-0.004	-0.001	-0.005
0.022545455	-0.001	0.009	-0.011
0.030893939	0.003	0.006	-0.001
0.019424242	0.014	0.019	0.017
0.029121212	0.011	0.013	0.005
0.008681818	0	0.002	0.003
0.014484848	0.002	0.004	-0.001
0.019939394	-0.021	-0.006	-0.019

데이터 전처리 오류

- large_data를 cmm_data.csv로 변경할 시 값이 없는 행이 발생

33	0.018015	0.005	0.015	0.004	0.008
34	0.02503	-0.005	0	-0.006	-0.004
35	0.032652	0.001	0.005	0.003	0.003
36					
37	0.008576	-0.016	-0.004	-0.012	-0.011
38	-0.01091	-0.003	0	0.006	0.001

- 다른 평가 항목으로 측정하기 때문인 것으로 추정

거리4 <-	원2(I) <상>, D	원3(I) <중>, D	원4(I) <하>, D	원통1
0.005				
0.005				
0.021				

Logistic Regression

1) Linear Combination 계산

- 입력 특성들과 각 특성에 대응하는 가중치들을 곱한 값을 모두 합하여 선형 결합(Linear Combination)을 계산

$$z = b + x_1 w_1 + x_2 w_2 + \cdots + x_n w_n$$

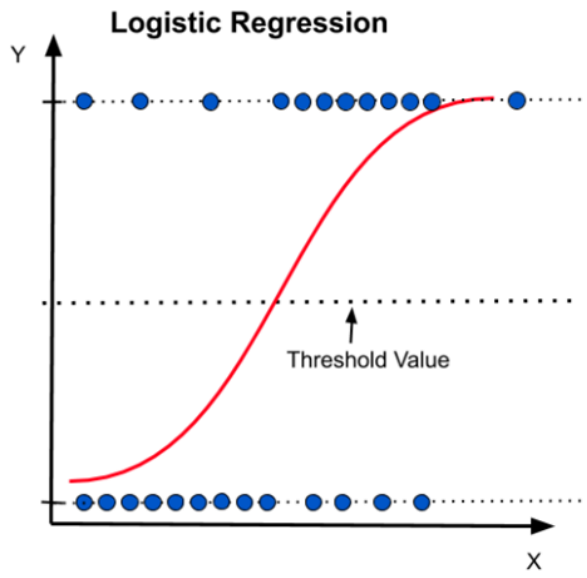
(w = weight, b = bias)

2) Logit을 확률로 변환

- 선형 결합 값을 Logit 함수의 역함수인 Sigmoid 함수에 통과시켜서, 0과 1 사이의 확률값 계산

3) 분류 결정

- 얻어진 확률값 p를 기준으로 분류 결정
- p가 0.5 이상이면 OK Class(1)로, 그렇지 않으면 NG Class(0)로 분류



Logistic Regression 예측 결과

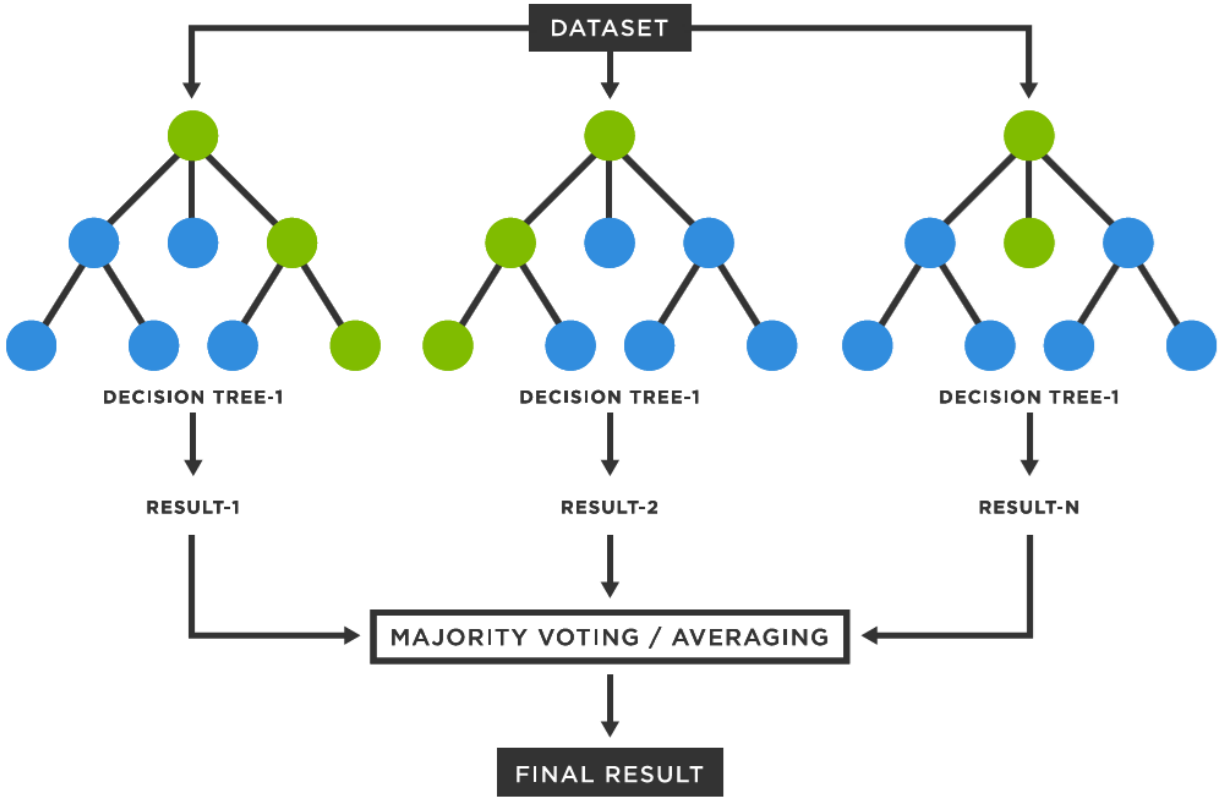
- 입력: 품질 상태를 제외한 각 도형, 항목에 따른 편차 데이터 (Train data: 110개 /Test data(support): 27개)
- 출력:
  - Test data에 따른 precision, recall, f1-score 값
  - 정확도: 약 48%
  - Macro avg: 약 42% (각 class 별 f1-score를 산술 평균한 값)
  - Weighted avg: 약 42% (각 class 별 f1-score를 support 수에 따라 가중치를 주어 평균한 값)

- small_dataset의 경우 정확도가 100%
  (Train data: 7개, Test data: 2개)

	precision	recall	f1-score	support
0	0.50	0.79	0.61	14
1	0.40	0.15	0.22	13
accuracy			0.48	27
macro avg	0.45	0.47	0.42	27
weighted avg	0.45	0.48	0.42	27

Random Forest

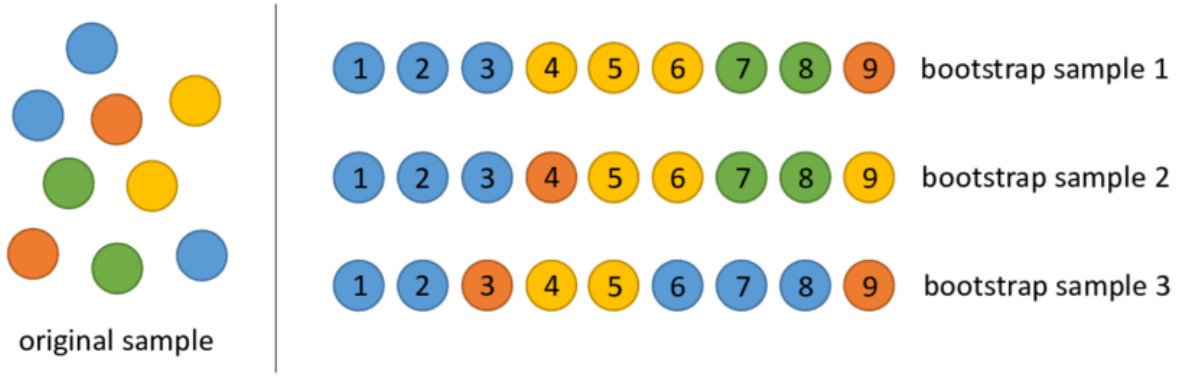
- Random Forest: Decision trees를 여러 개 만들어 그들의 예측을 평균 내는 앙상블 학습 모델
  - 앙상블 학습: 강력한 하나의 모델을 사용하는 대신 약한 여러 개의 모델을 조합하여 더 정확한 예측을 하는 방법



Random Forest 동작 과정

1) Training set에서 표본 크기가 n인 bootstrap sampling 수행

- bootstrap sampling: 원래 sample 집단에서 더 작지만 무수히 많은 집단으로 랜덤하게 뽑는 방법



2) Bootstrap sample에 대해 Random Forest Tree 모형 제작

- 전체 변수 중에서 m개 변수를 랜덤하게 선택
- 최적의 classifier 선정
- Classifier에 따른 2개의 node 생성

3) Tree들의 앙상블 학습 결과 출력

Random Forest 특성 분석 결과

- Feature Importance
  - 모델이 어떤 특성을 주로 사용하는지를 나타내는 지표
  - 각 특성이 모델의 예측에 얼마나 큰 영향을 미치는지를 평가
  - 모델이 어떤 특성을 분할 기준으로 선택하는지를 이해하는 데 도움

```
Top 5 Feature Importance:
Feature 원1(I) <상>, D: 0.08567732953370288
Feature 원통1(I) <- 원1, 원2, 원3의 측정점 병합, D: 0.07411592652787918
Feature 원3(I) <하>, D: 0.07046940093694092
Feature 원6(I) <하부>, Y: 0.03493388227673704
Feature 평면2, Z: 0.031270703305066545

Bottom 5 Feature Importance:
Feature 점28 <- 점27의 되부름 <소재 원점>, Y: 0.002788870262536898
Feature 점5 <- 직선8와 직선7의 교차점, Y: 0.003181073967927135
Feature 직선21 <우하 소재>, X/Y: 0.003719796447543111
Feature 거리1 <- XAXIS[PT]:점9와 점10 <상>, DS: 0.004166980448602631
Feature 거리4 <- XAXIS[평균]:점32와 점31 <소재기준>, DS: 0.004205369309733444
```

Random Forest 특성 분석 결과

- Permutation Importance
  - 실제 예측에 영향을 미치는 정도를 측정
  - 해당 특성을 무작위로 섞었을 때 모델의 성능이 얼마나 감소하는지를 측정함으로써 이루어짐
  - 각 특성이 모델의 예측에 실제로 얼마나 기여하는지를 보다 정확하게 파악하는 데 도움

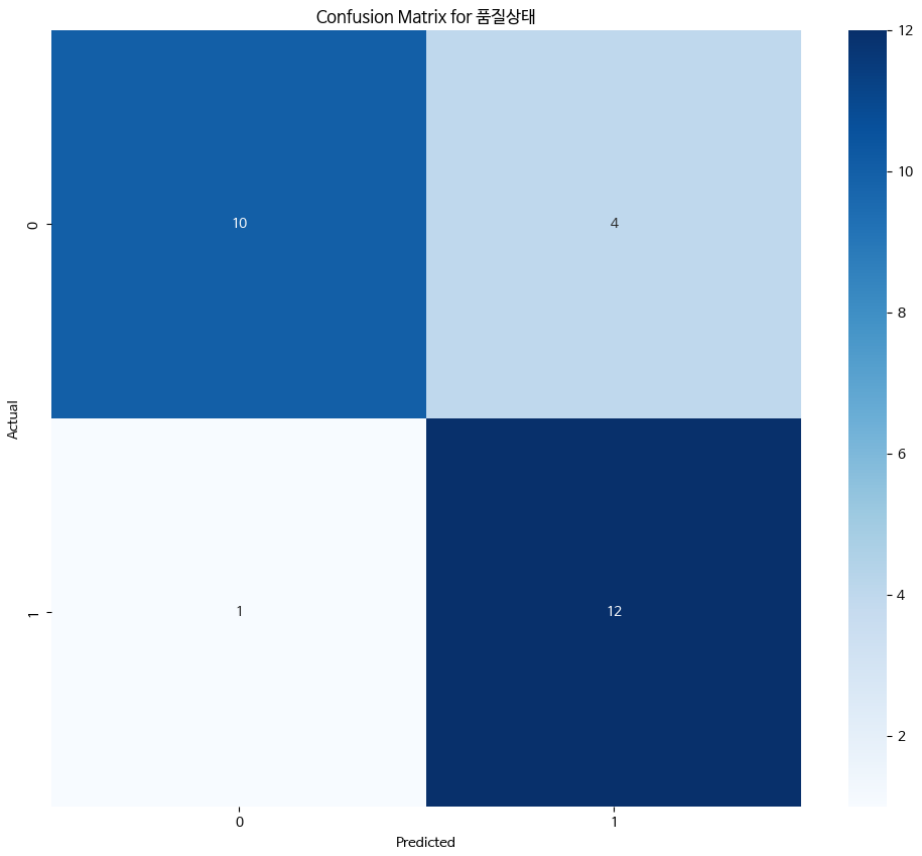
```
Top 5 Permutation Importance:
Feature 원4(E) <소재>, X: 0.08024691358024687
Feature 원2(I) <중>, D: 0.07901234567901232
Feature 점18 <- 점16와 점17의 중점 <열전관리_하>, X: 0.0716049382716049
Feature 점30 <- 점18의 되부름 <소재원점>, X: 0.06913580246913577
Feature 원7(E) <- 원4의 되부름, D: 0.06296296296296293

Bottom 5 Permutation Importance:
Feature 점20 <하>, Y: -0.016049382716049408
Feature 거리3 <- XYPLAN[PT]:원5와 원통1, DS: -0.016049382716049405
Feature 직선7 <23.1° >, Y/X: -0.014814814814814847
Feature 점6 <- 직선9와 직선10의 교차점 <소재>, X: -0.01111111111111113
Feature 직선25 <- 직선18의 되부름, X/Y: -0.008641975308642002
```

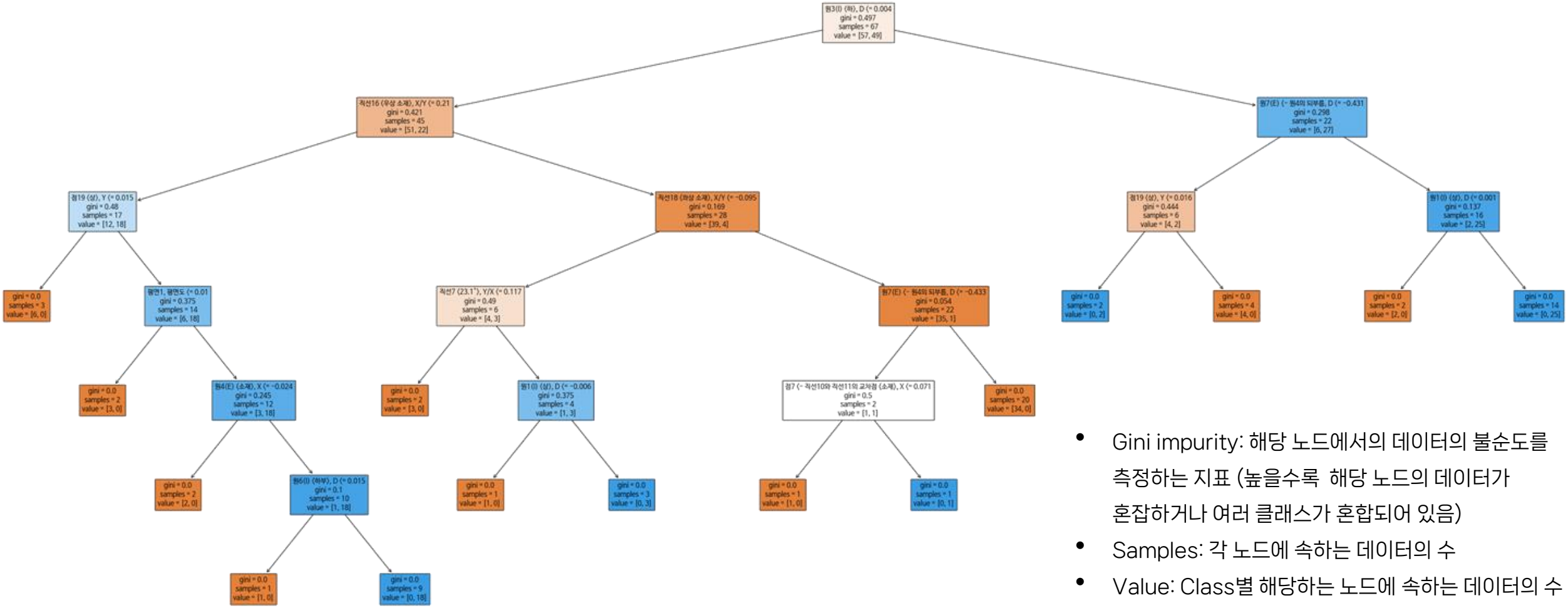
Random Forest 예측 결과

- 입력: 품질 상태를 제외한 각 도형, 항목에 따른 편차 데이터 (Train data: 110개 /Test data(support): 27개)
- 출력:
  - Test data에 따른 precision, recall, f1-score 값
  - 정확도: 약 81%
  - Macro avg 및 weighted avg: 약 81%

	precision	recall	f1-score	support
0	0.91	0.71	0.80	14
1	0.75	0.92	0.83	13
accuracy			0.81	27
macro avg	0.83	0.82	0.81	27
weighted avg	0.83	0.81	0.81	27



Random Forest 시각화



- Gini impurity: 해당 노드에서의 데이터의 불순도를 측정하는 지표 (높을수록 해당 노드의 데이터가 혼잡하거나 여러 클래스가 혼합되어 있음)
- Samples: 각 노드에 속하는 데이터의 수
- Value: Class별 해당하는 노드에 속하는 데이터의 수 [Class 0 인 노드의 수, Class 1인 노드의 수]

실증적AI프로젝트 금주 활동계획 (8주차)

주제: CMM 데이터의 이상치 탐지 딥러닝 모듈 개발

<table>
<tr><td>금주
활동계획</td><td colspan="3">1. CMM 데이터에 더 다양한 ML 모델 실험
2. GNN을 활용한 이상치 탐지 논문 리딩 및 발표</td></tr>
<tr><td></td><td>팀장 (김지선)</td><td>팀원1 (김예령)</td><td>팀원2 (백수민)</td></tr>
<tr><td>금주
개인별
활동계획</td><td>1. GNN 논문1 리딩 및 정리
• ML에 적용</td><td>2. GNN 논문2 리딩 및 정리
• G</td><td>3. 더 다양한 ML 모델 실험
• Timeseries_forcasting 계열 모델 적용
• Clustering 모델 적용</td></tr>
<tr><td>차주
활동계획</td><td colspan="3">1. GNN(Graph Neural Network) 이상치처리 모델 공부
2. GNN(Graph Neural Network) 이상치처리 모델 구현 실습</td></tr>
</table>

Thank you for Watching