**Preparation**

**[1]**

- 중복제거 및 유사한 행 필터링 : 일부 행은 중복되지만 날씨정보(온도/습도)만 약간의 차이가 있어 노이즈 제거 및 일관성 유지를 위해 날씨정보가 +-2 차이이면 유사하다고 보고, 그 중에서 대표값(결측치가 적은 행)을 선택하도록 했다.

- temperature\_C : 주방 온도에서 너무 높거나(60도) 너무 낮은(-10도) 이상치를 10도 미만 40도 초과이거나 Nan이면 최빈값으로 대체하도록 하였다. 온도는 이상치에 민감해 이상치의 영향을 덜 받는 최빈값을 선택하였다.

-meals\_served 결측치 제거 : meals\_served는 일부가 식사 제공량 정상 범위보다 10배 정도 극단적으로 높아서 (다 3자릿수인데 일부 4자릿수) 상위 1% 초과하는 행을 삭제하였다.

-past\_waste\_kg : past\_waste\_kg는 정상적으로 분포되어 있어 전체 경향을 잘 반영햐는 평균을 선택하였다.

-staff\_experience 결측치 처리 : kitchen\_staff와 past\_waste\_kg을 고려해 주방 인원 적고 쓰레기 많으면 beginner, 주방 인원 많고 쓰레기 적으면 expert, 나머지는 intermediate로 처리했다.

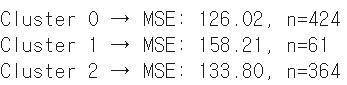
- 범주형 데이터 : staff\_experience는 경험 수준의 순서가 있으며 이후에 Regeression에서 숫자의 크기 자체에 의미를 부여하기 때문에 begginer:0, intermediate:1, exper:2로 직접 바꾸고 waste\_category는 알파벳 순으로 Label Encoding 하였다.

- env\_features = ['temperature\_C', 'humidity\_percent'] score가 높으면 Cluster 0

ops\_features = ['meals\_served', 'special\_event', 'day\_of\_week'] score가 높으면 Cluster 1

human\_features = ['kitchen\_staff', 'staff\_experience'] score가 높으면 Cluster 2

- 환경(env), 운영(ops), 인적(human) 요인을 기반으로 정규화(MinMax사용)된 점수를 만든 뒤, 각 샘플을 가장 높은 점수의 요인에 따라 클러스터로 분류하고, 그 클러스터별로 선형 회귀 모델을 학습해 성능(MSE)을 측정



-> Cluster 별 데이터 개수 차이가 큼

**[2]**

- temperature\_C값이 10~40 사이인 값들 중 최빈값(mode)으로 결측치를 채우고, 범위 밖의 값은 잘라냄(clip)

- past\_waste\_kg 결측치는 평균값으로 채움

- staff\_experience 결측치 규칙 기반 보정, 숫자형으로 매핑

-temperature\_C, humidity\_percent, meals\_served, kitchen\_staff네 개의 수치형 특성만 따로 지정한 뒤, SimpleImputer를 이용해 평균값으로 결측치를 채워줍니다.

- 타겟 변수인 food\_waste\_kg에 결측치가 있는 행들을 제거

- 전체 특성 값을 MinMaxScaler로 0과 1 사이로 정규화합니다. 이렇게 하면 값의 범위가 다르더라도 모델이 균형 있게 학습할 수 있게 됩니다.

- 정규화된 데이터를 기반으로 KMeans클러스터링을 수행하여 데이터를 세 개의 클러스터로 나눕니다. 클러스터링은 서로 다른 특성을 가진 그룹별로 모델을 따로 학습시키기 위해 사용됩니다. 클러스터 결과는 df['cluster']열에 저장

- Cluster 0. 온도/습도 낮고, 경험 적은 인원이 많은 경우

Cluster 1. 이벤트 없음, 식사수 많고 인원 충분한 패턴

Cluster2. 초보자 인력이 많은 특수한 날

- 클러스터에 대해 개별적으로 회귀 모델을 학습 -> 클러스터 번호를 기준으로 데이터를 필터링

-> 예측에 사용할 입력값(X)과 타겟값(y)을 분리한 후, 학습용 데이터와 검증용 데이터로 나눔

-> 모델은 RandomForestRegressor를 사용 -> 학습이 완료된 모델은 검증 데이터로 예측을 수행하고, 그 예측값과 실제값을 비교하여 평균 제곱 오차(MSE)를 계산

