2020 날씨 빅데이터 콘테스트

기상 데이터와 머신러닝, 딥러닝을 활용한

현대제철 결로 발생 예측 모델 제안

K 강상규

K 구본아

S 신우석

목차

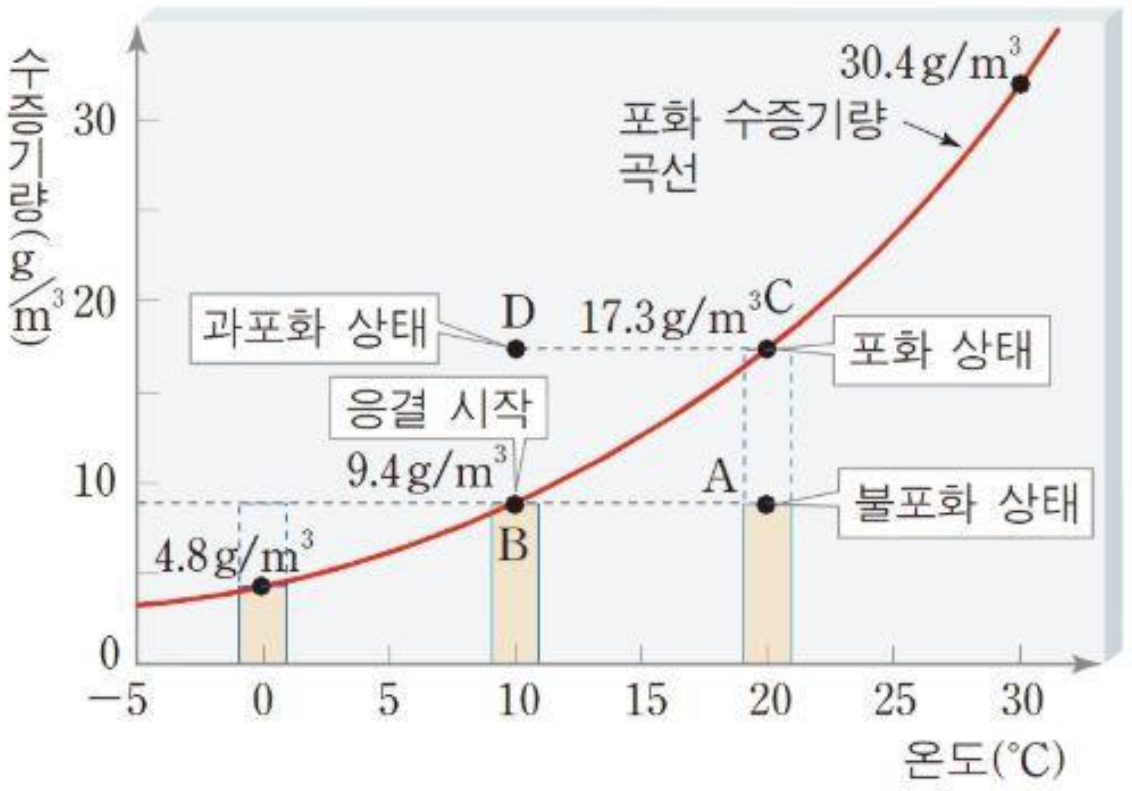
1. 문제정의

결로 발생을 예측하는 모델 …

결로가 무엇이길래 예측하려는 걸까?

2. 결로란?

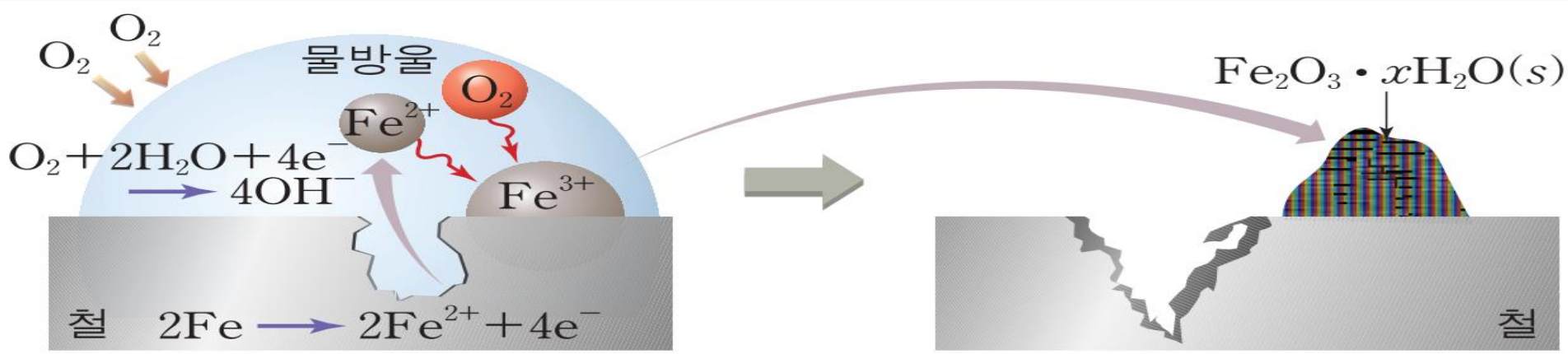
결로란,
내부 온도가 **이슬점 이하**로 떨어져 물체 표면에 **공기 중의 수증기**가 물방울로 맺히는 현상.



공기가 포함한 **수증기량**에 따라
온도가 변할때 **수증기에서 물로 변하는 지점**이 발생.

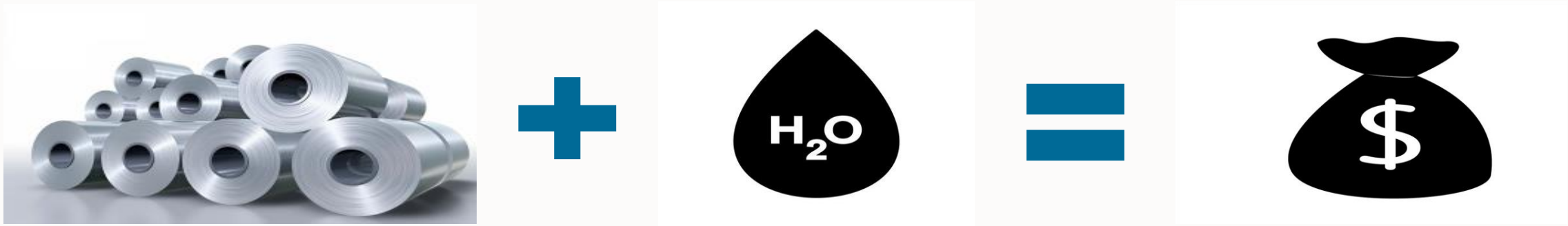
제철소는 **해안가**에 위치 해 있어 **온도, 습도**에 있어
결로가 발생하기 쉬운 **입지 조건**을 가짐.

3. 결로로 인한 비용발생



철강 기업이 해결해야할 문제 중 하나는 **철강제에 붙는 물방울**.

물방울이 철강제에 달라붙으면 **산화반응에 따른 녹**이 생성.



녹은 **철강제품의 품질을 하락**시키고, 철강기업은 그에 따른 **품질 비용**이 발생함.

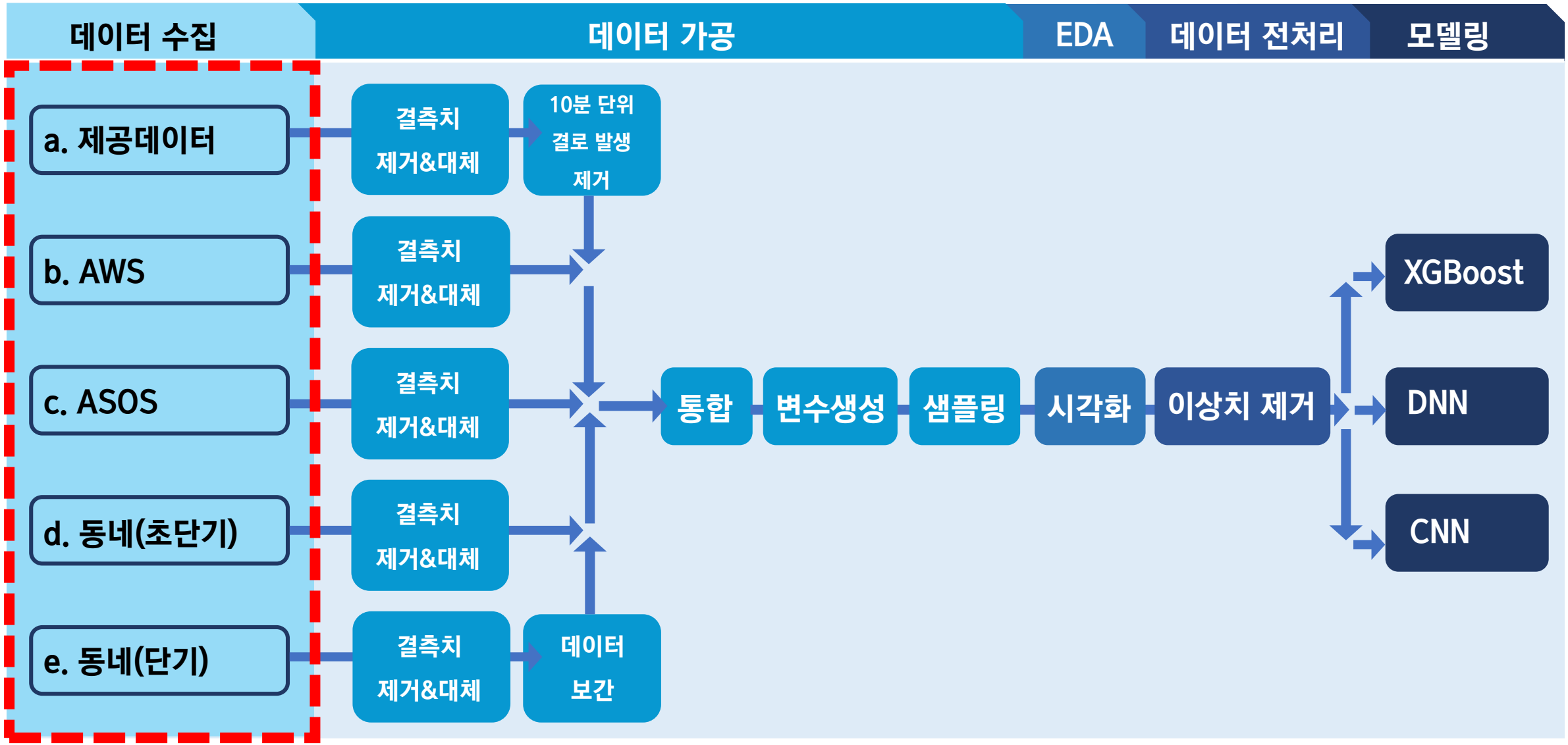
4. 목표

✓ 결로현상은 철강제품의 품질을 하락시키고 **경제적 손실** 초래.

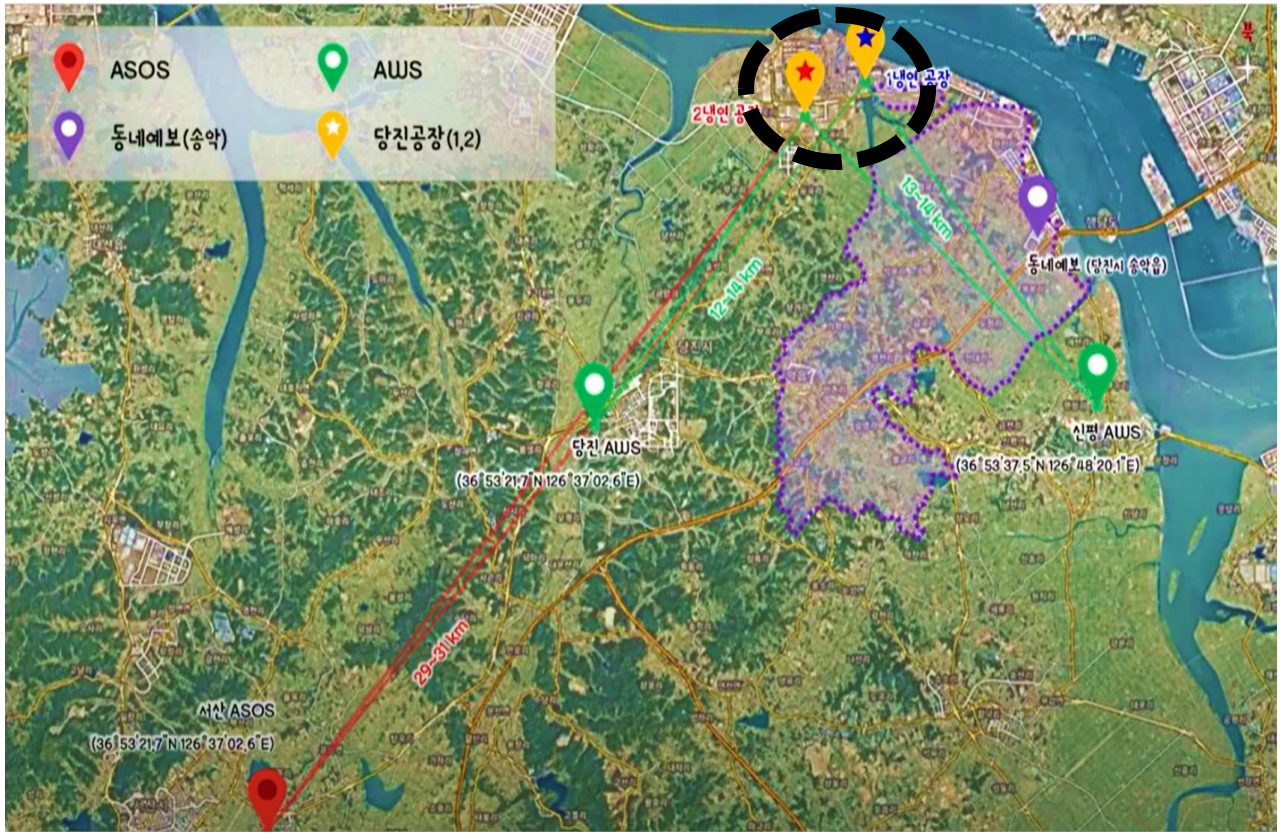
✓ 현대제철은 생산된 철강제품에 발생하는 **결로현상을 방지하는 방안** 모색.

✓ **기상 데이터**와 **머신러닝 및 딥러닝 기법**을 활용해

철강제품의 **결로 발생을 예측**함으로써, **손실 비용을 최소화**하는 것이 목표.


데이터 수집
데이터 가공
EDA
데이터 전처리
모델링
a. 제공데이터
b. AWS
c. ASOS
d. 동네(초단기)
e. 동네(단기)
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
10분 단위 결로 발생 제거
데이터 보간
통합
변수생성
샘플링
시각화
이상치 제거
XGBoost
DNN
CNN

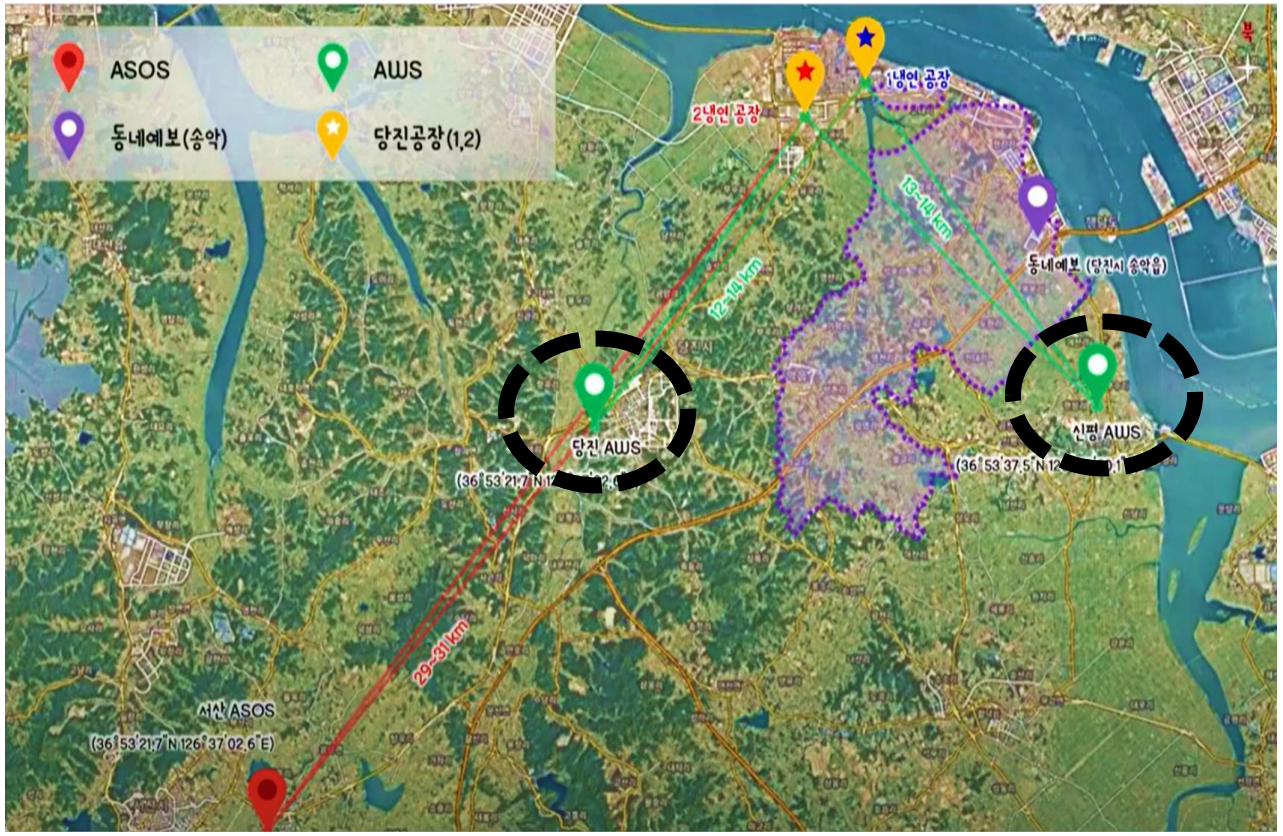
1. 제공 데이터 (현대제철)



위치	총 Data	변수	결측치
당진	340980	9	1094

변수속성	당진
날짜	plant_mea_ddhr
공장	plant
위치	loc
공장 내부 온도	plant_temp_in
공장 내부 습도	plant_hum_in
코일 온도	plant_tem_coil
공장 외부 온도	plant_temp_out
공장 외부 습도	plant_hum_out
24시간 후 결로 발생 여부	plant_cond_24
48시간 후 결로 발생 여부	plant_cond_48

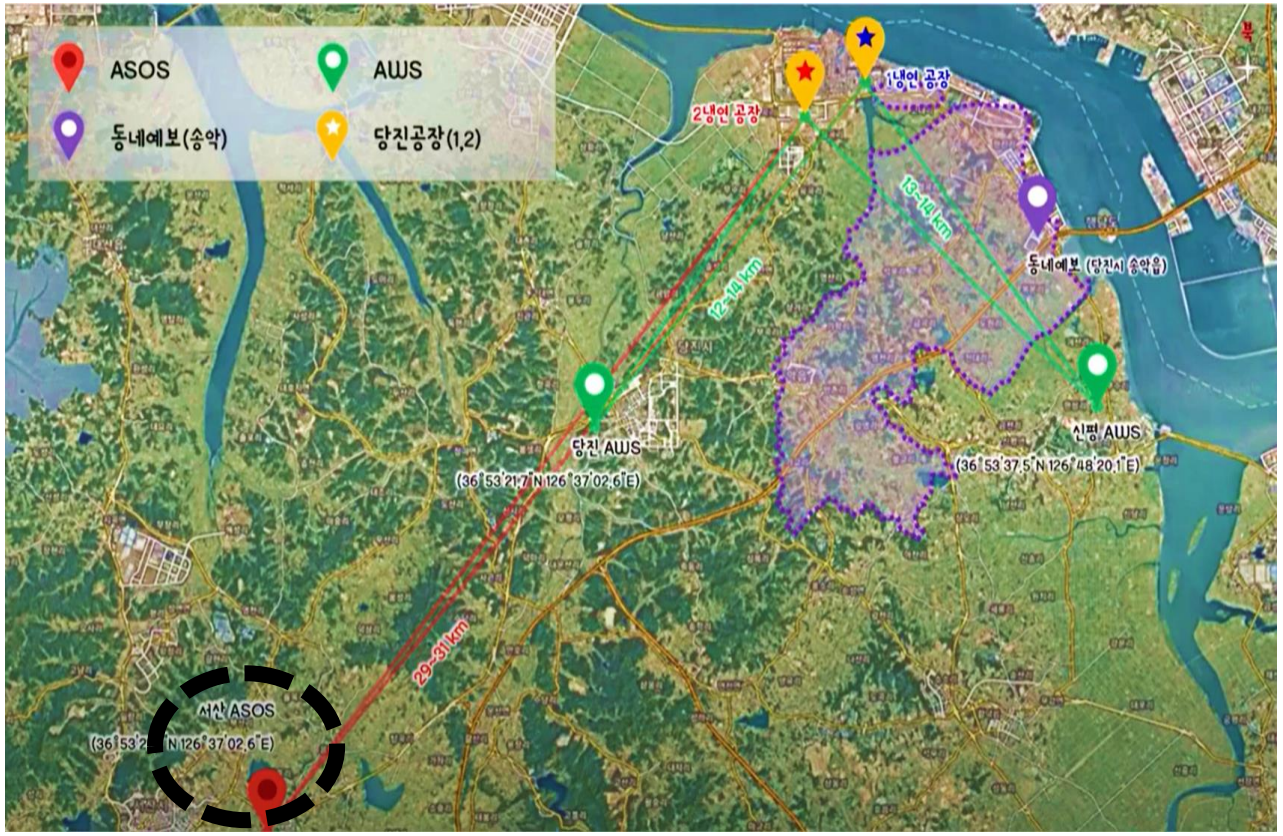
2. AWS



위치	총 Data	변수	결측치
당진	32248	5	317
신평	32337	6	256

변수속성	당진	신평
기온	temperature_dj	temperature_sp
습도	humidity_dj	-
풍향	wind_deg_dj	wind_deg_sp
풍속	wind_speed_dj	wind_speed_sp
강수량	rainfall_dj	rainfall_sp
지상기압	-	local_pressure_sp
해상기압	-	sea_pressure_sp

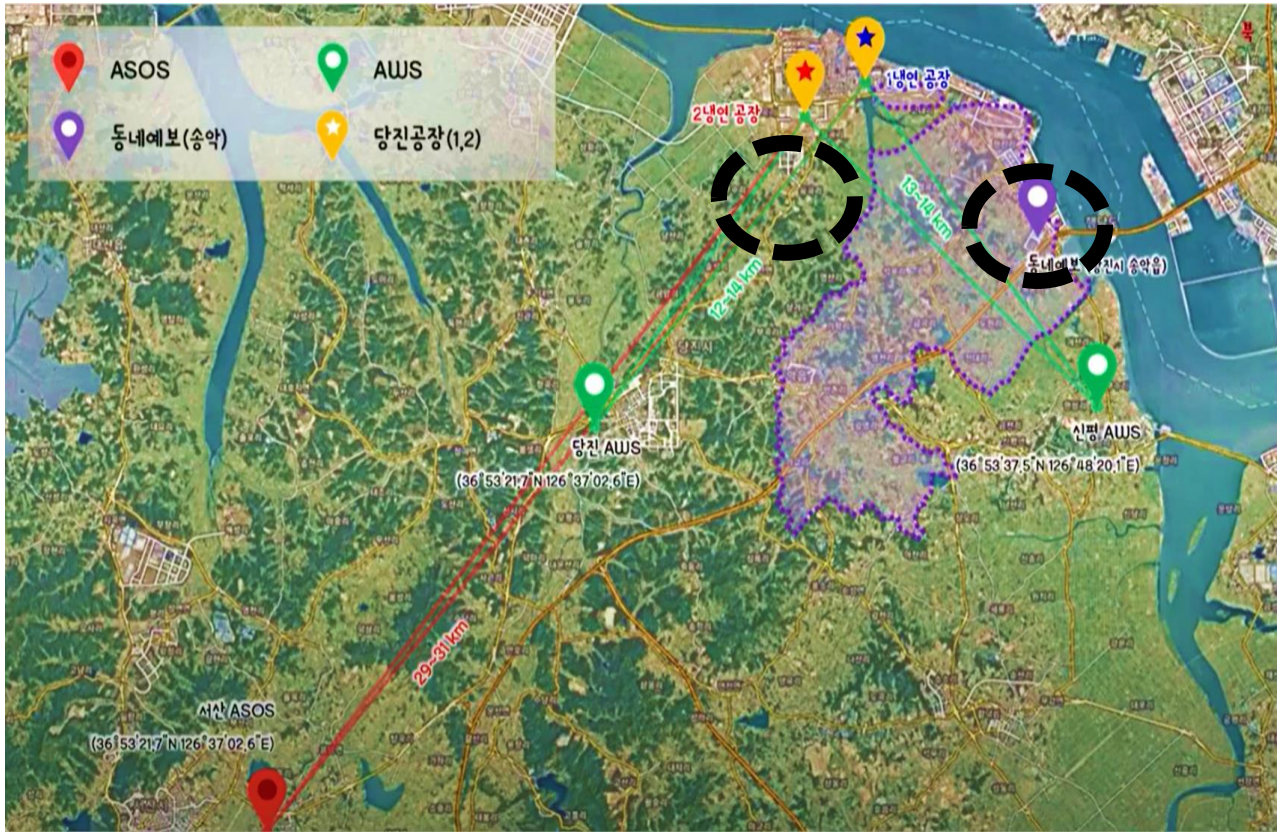
3. ASOS



위치	총 Data	변수	결측치
서산	32273	8	450

변수속성	서산
기온	temperature_ss
습도	humidity_ss
풍향	wind_deg_ss
풍속	wind_speed_ss
지상기압	ground_pressure_ss
해상기압	sea_pressure_ss
지상기온	ground_temp_ss
이슬점 온도	dewpoint_ss

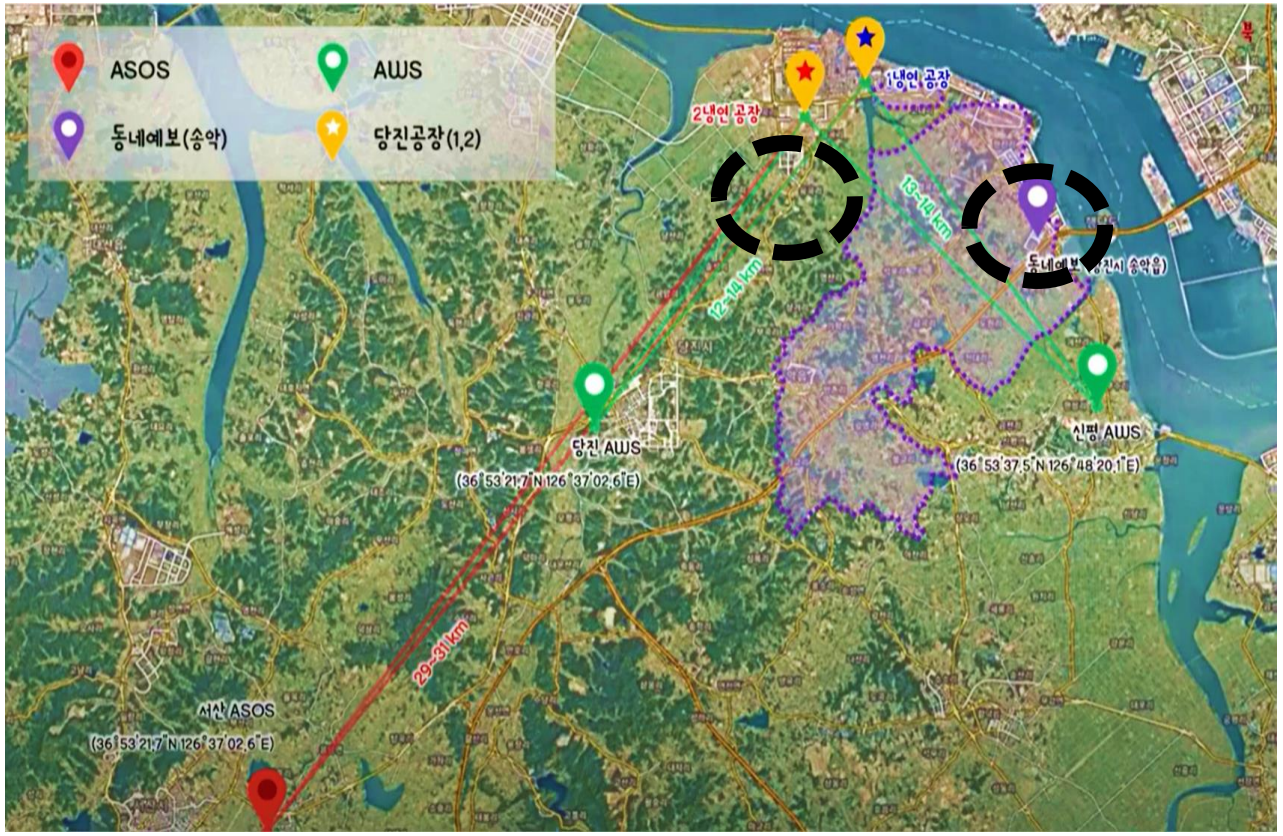
4. 동네 (초단기)



위치	총 Data	변수	결측치
송악읍	32880	6	0
송산면	32880	6	0

변수 속성	송악읍	송산면
기온	temperature_sa	temperature_ssm
습도	humidity_sa	humidity_ssm
풍향	wind_direc_sa	wind_direc_ssm
풍속	wind_speed_sa	wind_speed_ssm
강수량	rain_sa	rain_ssm
강수형태	rain_type_sa	rain_type_ssm

5. 동네 (단기)

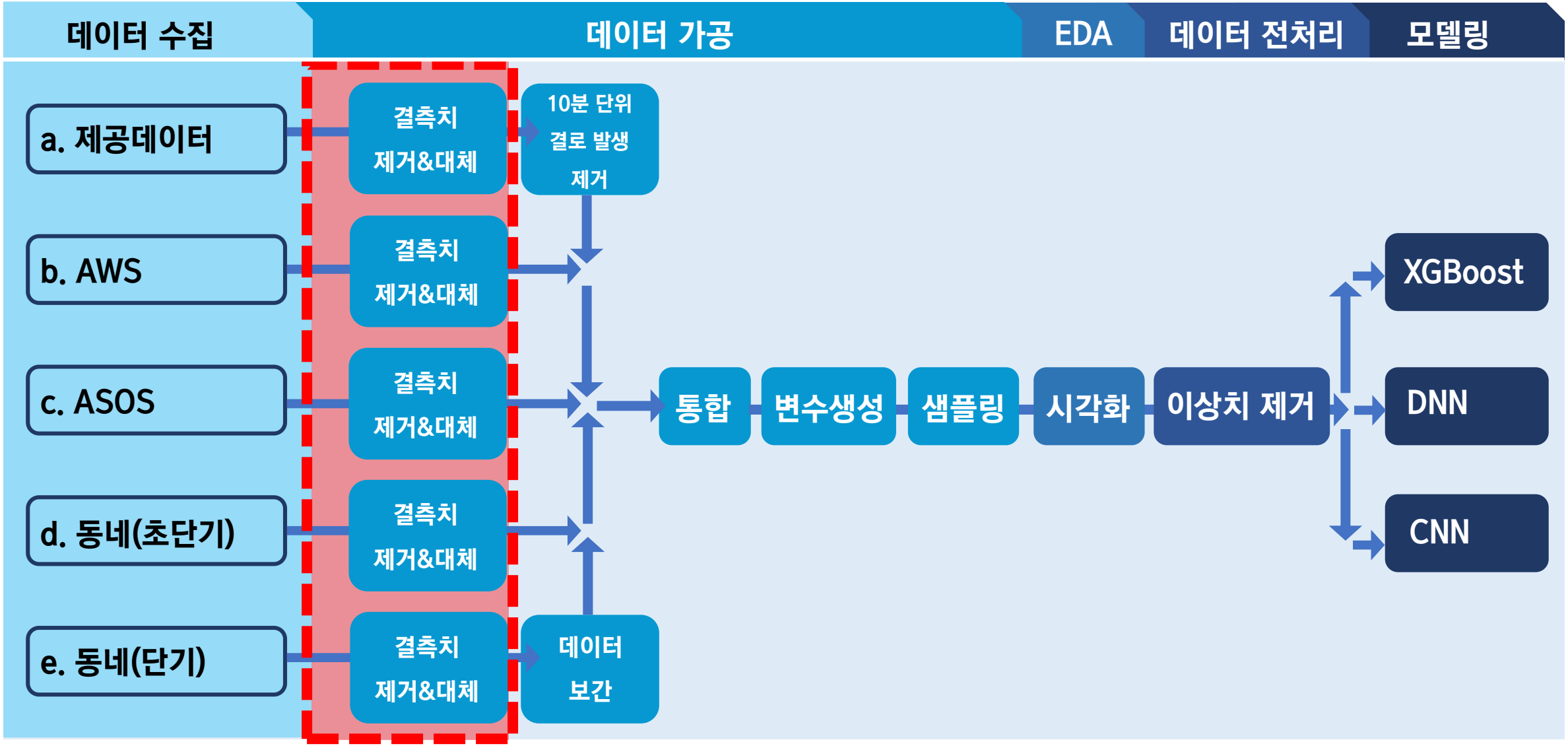


위치	총 Data	변수	결측치
송악읍	65755	42	1340
송산면	65755	42	1340

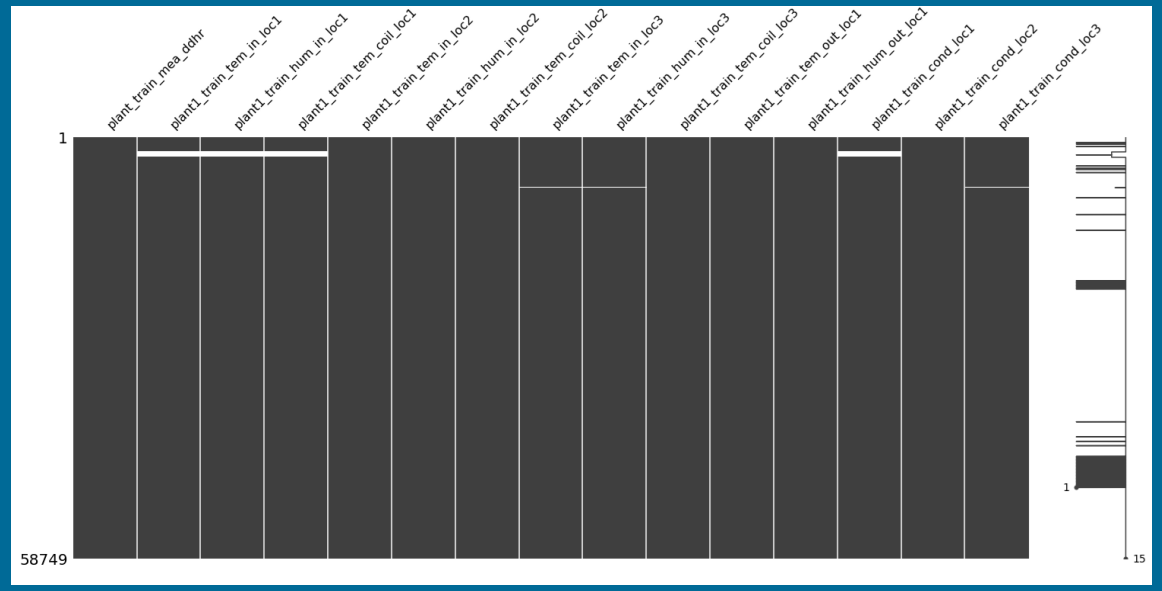
변수 속성	송악읍	송산면
3시간 기온 예측	3h_temp_pred_sa	3h_temp_pred_ssm
습도 예측	hum_pred_sa	hum_pred_ssm
풍향 예측	wind_direc_pred_sa	wind_direc_pred_ssm
풍속 예측	wind_speed_pred_sa	wind_speed_pred_ssm
강수확률 예측	rain_proba_pred_sa	rain_proba_pred_ssm
하늘상태 예측	sky_cond_pred_sa	sky_cond_pred_ssm

현대제철 데이터의 기준일

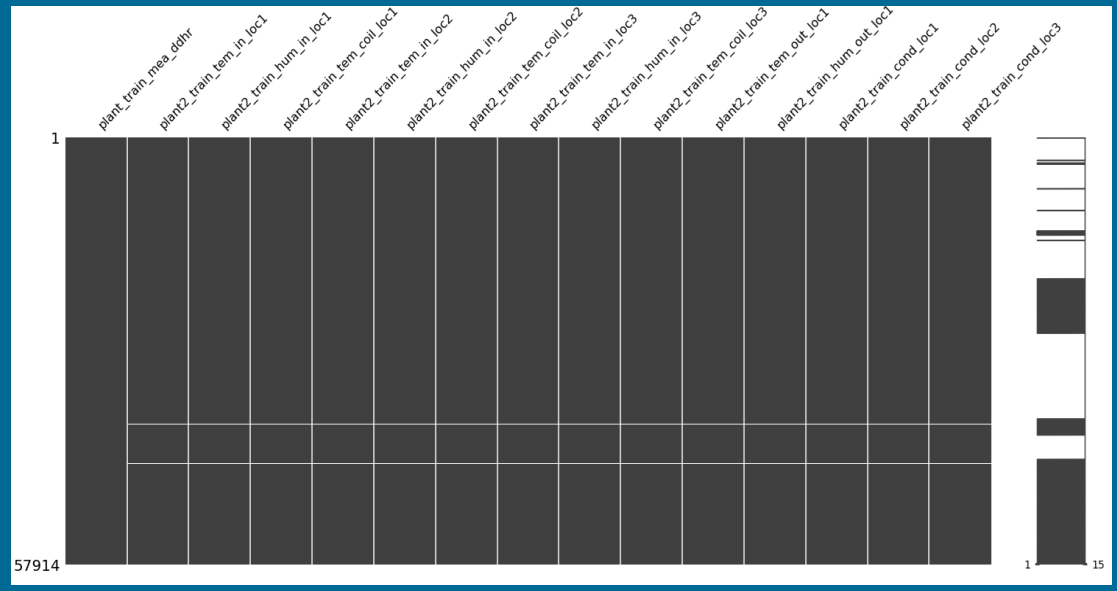
13, 19, 25, 31, 37, 43, 49 시간 후 예보 데이터 활용


데이터 수집
데이터 가공
EDA
데이터 전처리
모델링
a. 제공데이터
b. AWS
c. ASOS
d. 동네(초단기)
e. 동네(단기)
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
10분 단위 결로 발생 제거
데이터 보간
통합
변수생성
샘플링
시각화
이상치 제거
XGBoost
DNN
CNN

1. 결측치 처리



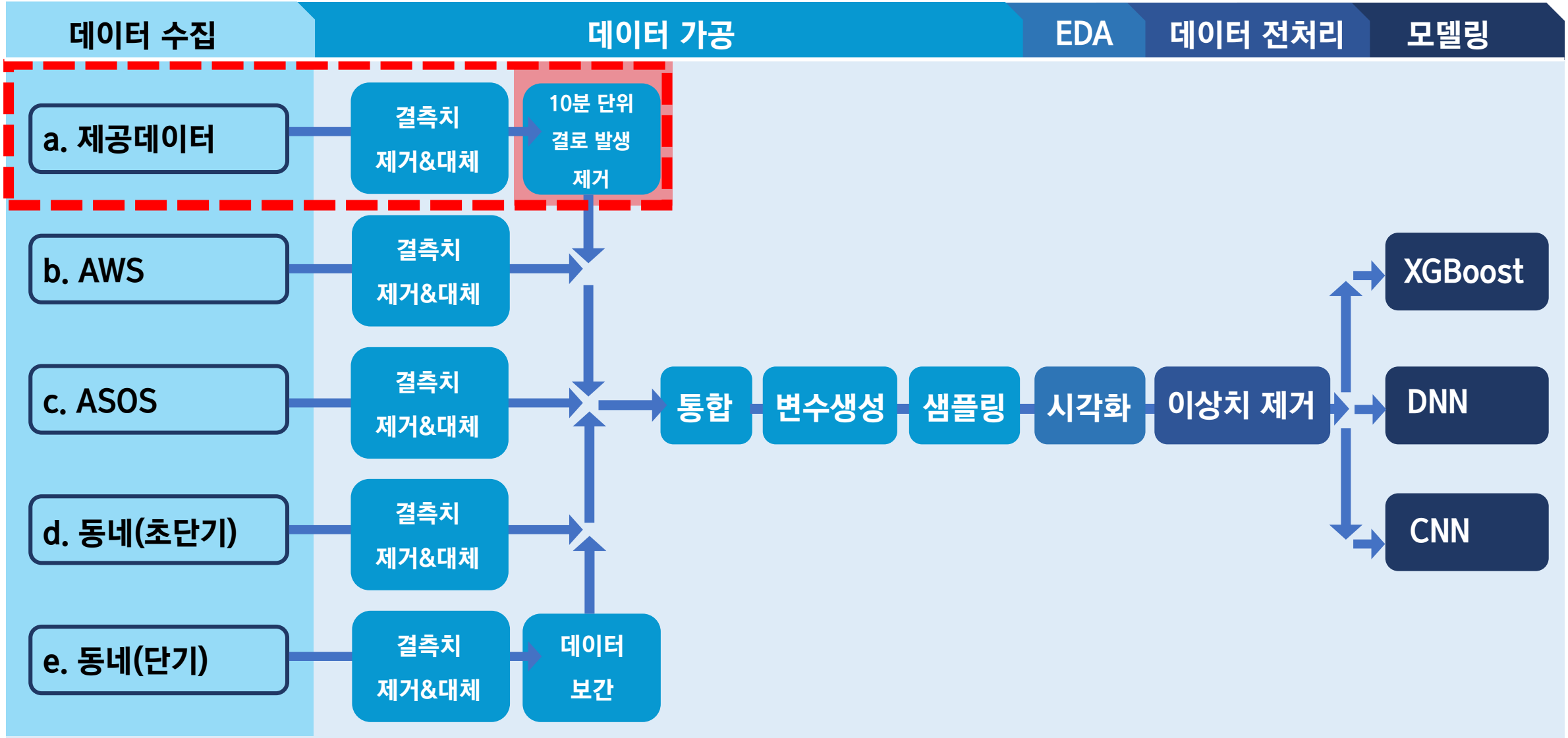
plant_1 missing value



plant_2 missing value

- ✓ 현대제철 데이터의 경우, 결측치에 패턴 존재.
- ✓ 측정 장비 정비로 인한 결측치로 판단하여 제거. (plant_1 870개, plant_2 224개)
- ✓ AWS, ASOS, 초단기예보, 단기예보 데이터의 결측치는 이전 값으로 대체.

2. 현대제철 데이터 10분 단위 결로 발생 제거



2. 현대제철 데이터 10분 단위 결로 발생 제거

날짜	습도	…	코일온도	외부온도
2016.07.20 : 12:00	30.2	…	10.1	20.3
….	…	…	…	…
2018.7.01 : 12:00	28.9	…	8.9	25.6
2018.7.01 : 12:10	28.1	…	9.5	24.5
…	…	…	…	…

현대제철 데이터

날짜	풍속	…	강수량	하늘상태
2016.07.20 : 12:00	43.7	…	0.0	1
….	…	…	…	…
2018.7.01 : 12:00	70.6	…	2.3	4
2018.7.01 : 13:00	88.4	…	1.1	3
…	…	…	…	…

외부 기상 데이터

✓ 병합하려는 외부 기상 데이터가 1시간, 3시간 단위로 기록.

✓ 그러나, 현대제철 데이터의 2018년 6월 이후는 10분 단위로 구성되어 처리가 필요.

✓ 따라서, 30분 단위와 10분 단위 데이터의 정보량을 비교.

2. 현대제철 데이터 10분 단위 결로 발생 제거

공장	위치	내부 온도	내부 습도	코일 온도	외부 온도	외부 습도
1	1	46.90	102.94	36.09	79.40	249.35
	2	49.68	123.80	41.20	68.99	230.96
	3	52.74	140.20	46.15	72.88	243.88
2	1	36.35	106.00	28.12	74.67	330.04
	2	56.34	151.36	43.57	90.47	382.46
	3	7.57	72.21	5.60	16.37	150.48

30분 단위 결로 데이터 분산(Variance)

VS

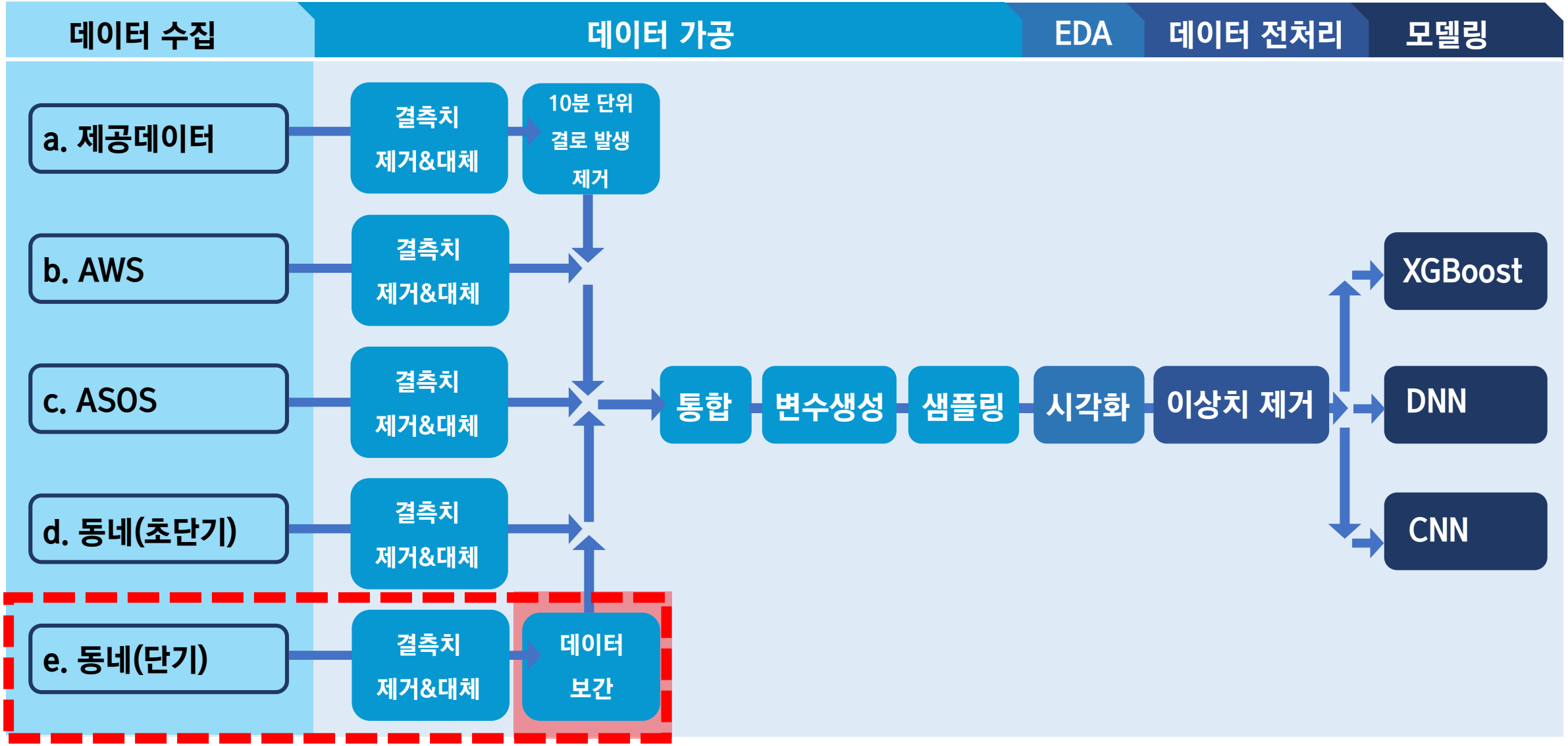
공장	위치	내부 온도	내부 습도	코일 온도	외부 온도	외부 습도
1	1	22.85	29.74	19.38	28.45	100.06
	2	32.63	55.90	26.86	41.38	121.17
	3	42.77	73.61	38.31	53.98	148.27
2	1	29.03	51.08	25.78	45.42	186.51
	2	33.75	104.28	28.63	46.98	238.75
	3	0.03	0.15	0.08	0.68	11.25

10분 단위 결로 데이터 분산(Variance)

✓ 30분 단위 결로 발생 데이터의 분산(정보량)과 10분 단위 결로 발생 데이터의 분산을 비교.

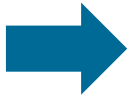
✓ 10분 단위 데이터의 분산은 30분 단위 데이터의 분산에 비해 매우 작았기에,
모델이 학습 할 때 10분 단위의 결로에 편향되어 학습될 수 있어 제거.

3. 예보 데이터 날짜 형식 변경 및 보간



3-1. 예보 데이터 날짜 형식 변경

day	hour	forecast	습도	…
1	200	4	70.1	
1	200	7	65.8	
…	…	…	…	
1	2300	67	54.3	
2	200	4	4	
…	…	…	…	



날짜	13시간 후 습도	19시간 후 습도	…	49시간 후 습도	All Features
2016.07.1 : 2:00	40.2	30.5	45.5	66.7	25.1
2018.7.01 : 5:00	40.5	70.5	75.1	80.8	20.4
2018.7.01 : 8:00	56.9	60.5	8.9	80.1	17.4
…	…	…	…	…	…
2020.3.31 : 23:00	40.1	25.5	30.2	35.1	30.1

✓ 동네 예보 데이터는 왼쪽과 같은 형태로 3시간마다 하나의 특정 날짜, 시간에 '4시간 후 ~ 67시간 후' 까지 예보를 발표.

✓ 기존 데이터와 병합하기 위해 오른쪽과 같이 하나의 특정 날짜, 시간에 '13시간 후 ~ 49시간 후' 까지 예보 변수를 6시간 단위로 처리.

✓ 매일 17:00시는 49시간 후의 예보가 존재하지 않음 → 46시간 후의 예보로 보간 (1340개의 행)

3-2. 예보 데이터 보간

날짜	13시간 후 습도	19시간 후 습도	…	49시간 후 습도	All Features
2016.7.01 : 2:00	40.2	30.5	45.5	66.7	25.1
2018.7.01 : 5:00	40.5	70.5	75.1	80.8	20.4
2018.7.01 : 8:00	56.9	60.5	8.9	80.1	17.4
…	…	…	…	…	…
2020.3.31 : 23:00	40.1	25.5	30.2	35.1	30.1

✓ **3시간 단위의 동네예보 데이터를**

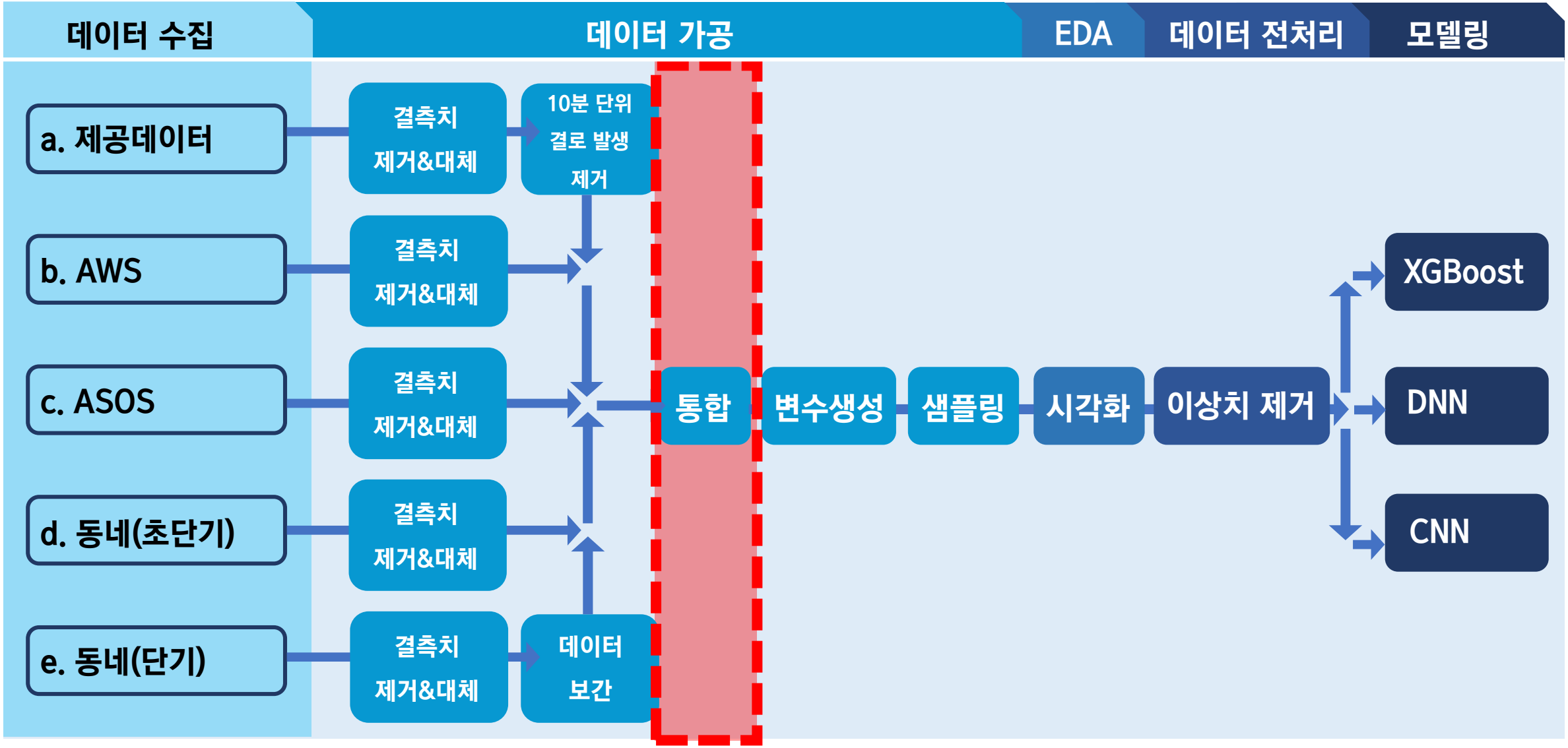
30분 단위의 테스트 셋에 적용하기 위해 30분 단위로 보간.

✓ **기후 데이터는**

시계열적인 특성을 가지므로 가중치를 주어 보간을 진행.

날짜	13시간 후 습도	19시간 후 습도	…	49시간 후 습도	All Features
2018.7.01 : 2:30	40.025	37.18	50.44	69.05	24.32
2018.7.01 : 3:00	40.30	43.82	55.36	71.40	23.53
2018.7.01 : 3:30	40.35	50.50	60.30	73.75	22.75
2018.7.01 : 4:00	40.40	57.18	65.24	76.10	21.97
2020.7.01 : 4:30	40.45	63.82	70.16	78.44	21.18

보간 계산 방법
7월 1일 2시 30분 = (0.833 x 7월 1일 2시) + (0.167 x 7월 1일 5시)
7월 1일 3시 00분 = (0.667 x 7월 1일 2시) + (0.333 x 7월 1일 5시)
7월 1일 3시 30분 = (0.5 x 7월 1일 2시) + (0.5 x 7월 1일 5시)
7월 1일 4시 00분 = (0.333 x 7월 1일 2시) + (0.667 x 7월 1일 5시)
7월 1일 4시 30분 = (0.167 x 7월 1일 2시) + (0.833 x 7월 1일 5시)


데이터 수집
데이터 가공
EDA
데이터 전처리
모델링
a. 제공데이터
b. AWS
c. ASOS
d. 동네(초단기)
e. 동네(단기)
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
10분 단위 결로 발생 제거
데이터 보간
통합
변수생성
샘플링
시각화
이상치 제거
XGBoost
DNN
CNN

4-1. 현대제철 데이터 병합

현대제철 공장 1 데이터

날짜	공장1_위치1_온도	...	공장1_위치1_결로
2016.07.20 : 12:00	30.2	...	0
2016.07.20 : 15:00	33.4	...	0
2016.07.20 : 18:00	28.9	...	1
2016.07.20 : 21:00	28.1	...	0

현대제철 공장 2 데이터

날짜	공장2_위치1_온도	...	공장2_위치1_결로
2016.07.20 : 12:00	33.7	...	1
2016.07.20 : 15:00	31.2	...	0
2016.07.20 : 18:00	29.5	...	0
2016.07.20 : 21:00	30.9	...	0

날짜	공장	위치	온도	...	24시간 후 결로	48시간 후 결로
2016.07.20 : 12:00	1	1	30.5	...	0	0
2016.07.20 : 15:00	1	2	28.7	...	0	0
2016.07.20 : 18:00	1	3	31.3	...	0	1
...	...	...	...	...	...	...
2016.07.20 : 12:00	2	1	33.1	...	1	0
2016.07.20 : 15:00	2	2	29.8	...	0	1
2016.07.20 : 18:00	2	3	30.7	...	1	0

✓ 현대제철에서 제공한 공장1, 공장2 Train set을

Test set 형식에 맞게 변경.

✓ 기준일 + 24시간, 48시간 결로 발생 여부를 타겟 변수로 생성.

4-2. 기상 데이터 병합

현대제철 데이터

날짜	습도	...	코일온도	이슬점
2016.07.20 : 12:00	30.2	...	10.1	20.3
2016.07.20 : 15:00	33.4	...	9.7	22.7
2016.07.20 : 18:00	28.9	...	8.9	25.6
2016.07.20 : 21:00	28.1	...	9.5	24.5

AWS 데이터

날짜	풍속	...	강수량	기압
2016.07.20 : 12:00	43.7	...	0.0	60.5
2016.07.20 : 15:00	60.5	...	0.0	70.4
2016.07.20 : 18:00	70.6	...	2.3	67.5
2016.07.20 : 21:00	88.4	...	1.1	71.2

ASOS 데이터

날짜	강수형태	...	온도	습도
2016.07.20 : 12:00	1	...	30.5	60.2
2016.07.20 : 15:00	0	...	28.7	57.3
2016.07.20 : 18:00	0	...	33.1	55.4
2016.07.20 : 21:00	2	...	29.8	58.7

동네예보 데이터

날짜	습도예측	...	강수확률	하늘상태
2016.07.20 : 12:00	30.8	...	0.0	1
2016.07.20 : 15:00	27.2	...	10.0	2
2016.07.20 : 18:00	33.9	...	30.0	3
2016.07.20 : 21:00	29.7	...	0.0	1

현대제철 데이터,

AWS 데이터,

ASOS 데이터,

동네예보 데이터를

날짜와 시간을 기준으로 병합

(30분 단위)

4-3. 데이터 병합

	현대제철 데이터						ASOS 데이터		AWS 데이터				동네예보 데이터			
날짜	습도	…	코일 온도	외부 온도	강수 형태	…	온도	습도	풍속	…	강수 량	기압	예측 습도	…	하늘 상태	강수 확률
2016.07.20 : 12:00	30.2	…	10.1	20.3	1	…	30.5	60.2	43.7	…	0	60.5	30.8	…	1	0.05
2016.07.20 : 15:00	33.4	…	9.7	22.7	0	…	28.7	57.3	60.5	…	0	70.4	27.2	…	3	0.2
2016.07.20 : 18:00	28.9	…	8.9	25.6	0	…	33.1	55.4	70.6	…	2.3	67.5	33.9	…	3	0.3
2016.07.20 : 21:00	28.1	…	9.5	24.5	2	…	29.8	58.7	88.4	…	1.1	71.2	29.7	…	4	0.6

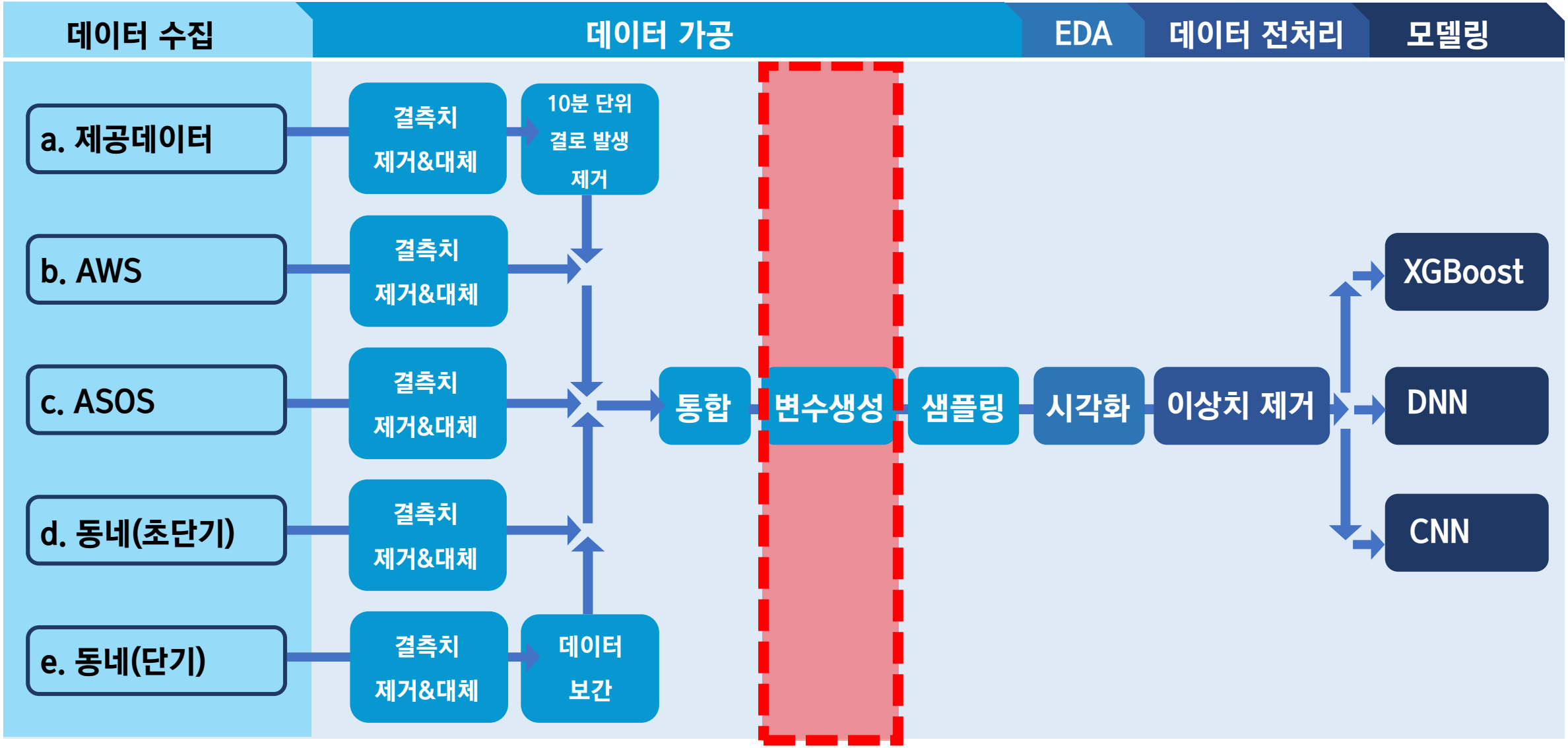
✓ 현대제철 데이터 : 7개 변수

✓ ASOS 데이터 (서산) : 8개 변수

✓ AWS 데이터 (당진, 신평) : 5개 + 6개 = 총 11개 변수

✓ 동네예보-초단기실황, 단기예보 데이터 (송악읍, 송산면) : 42 + 42 = 총 84개 변수

병합 후 약 16만개의 Row, 120개 Feature


데이터 수집
데이터 가공
EDA
데이터 전처리
모델링
a. 제공데이터
b. AWS
c. ASOS
d. 동네(초단기)
e. 동네(단기)
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
10분 단위 결로 발생 제거
데이터 보간
통합
변수생성
샘플링
시각화
이상치 제거
XGBoost
DNN
CNN

5. 변수 생성

✓ 결로는 계절, 온도, 습도 등 기상 데이터의 차이 및 비율에 의해 발생.

✓ EDA 결과를 토대로 변수 생성 (총 500여개 변수를 학습에 활용)

1. 현대제철 데이터

1) 월별로 결로수의 차이가 심함 → Month 변수 생성

2) 겨울(11월, 12월, 1월, 2월), 봄(3월, 5월)에 상대적으로 결로가 많이 발생함.

→ 겨울(11월, 12월, 1월, 2월), 봄(3월, 4월, 5월), 여름(6월, 7월, 8월, 9월, 10월) season 변수 생성

3) 내부 온 · 습도와 코일 온도, 외부 온 · 습도 변수들의 조합으로 비율(Ratio) 변수 생성

e.g.) 내부 온도 / 내부 습도, 내부 온도 / 외부 온도 · · ·

2. 단기예보 데이터

1) 결로는 온도차와 습도가 중요 → 각 예보 변수들의 예보 시점 간 차이 변수 생성 (e.g. 25시 예보 온도 – 19시 예보 온도)

2) 이슬점 변수 생성

3. 4가지 모든 데이터셋

1) 날짜별로 각 변수들의 최대값(Max), 최소값(Min), 최대값 – 최소값(Max – Min) 변수 생성

5. 변수 생성

- ✓ 결로는 내부 온도가 이슬점 이하로 떨어질 때 발생 → 이슬점과 관련되어 있음.
- ✓ 이슬점은 공기가 포화되어 수증기가 응결할 때의 온도를 말하거나, 불포화 상태의 공기가 냉각될 때 포화되어 응결이 시작되는 온도
- ✓ 즉, 온도가 낮아지면 공기가 수증기를 함유할 수 있는 양이 줄어들게 되고, 공기가 이슬점 이하로 냉각되어 포화상태가 되면 수증기가 물방울로 맺힘.
- ✓ 따라서, 실제로 결로가 발생하게 되는 해당 지점의 특성이 중요할 것으로 생각함.

➡ **현대제철 데이터와 단기예보 데이터의 온도와 습도 변수를 활용해 '이슬점' 변수 생성**

✓ **이슬점 생성 공식**

$$gamma = \frac{17.62 + temperature}{243.12 + temperature} + \ln(\frac{humidity}{100})$$

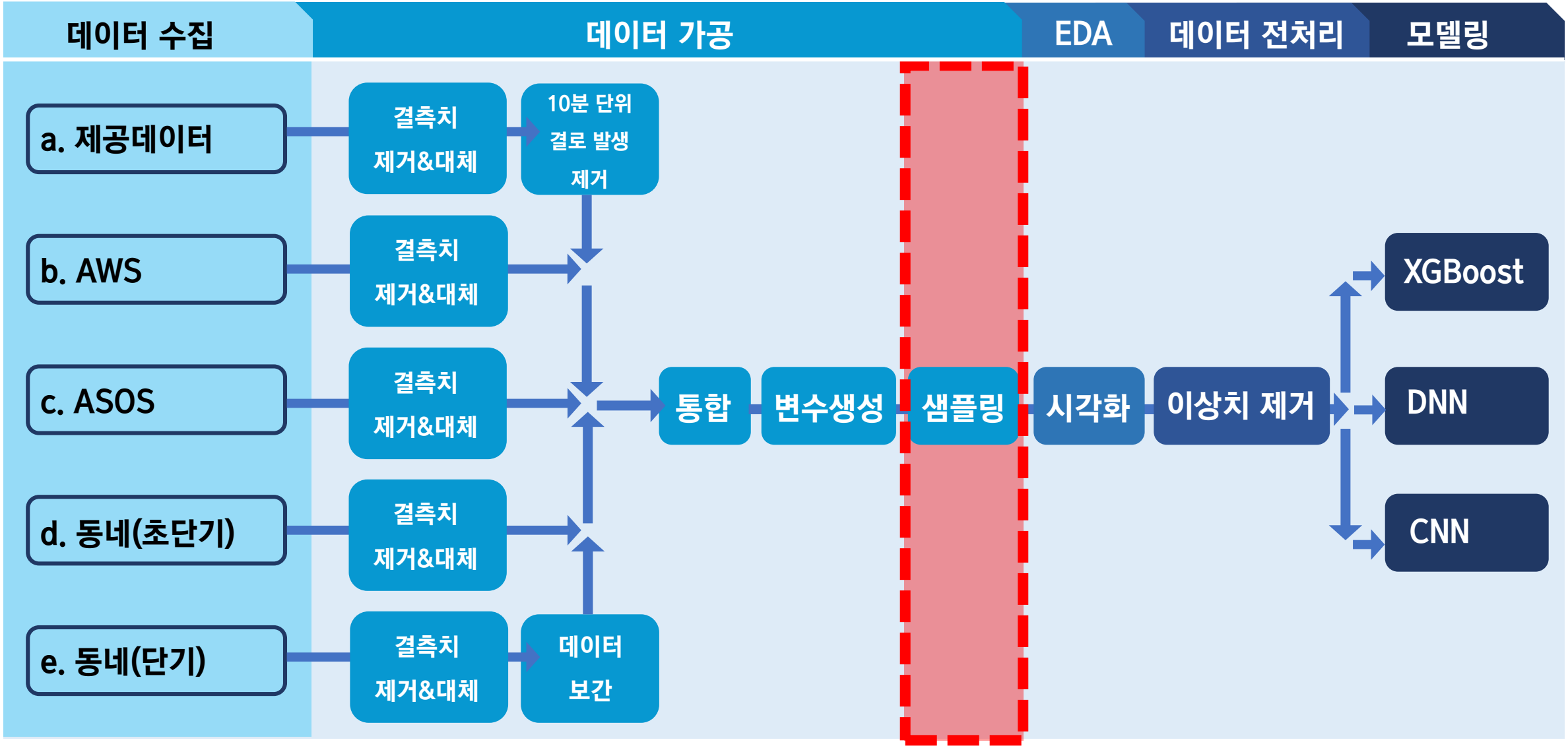
$$dewpoint = \frac{243.12 * gamma}{17.62 - gamma}$$

현대제철 데이터 : 세 가지의 이슬점 변수를 생성

(코일 온도, 내부 습도), (내부 온도, 내부 습도), (외부 온도, 외부 습도)

단기예보 데이터 : 예보 기온 평균과 습도를 사용해 이슬점 변수 생성

6. 샘플링(Under Sampling)

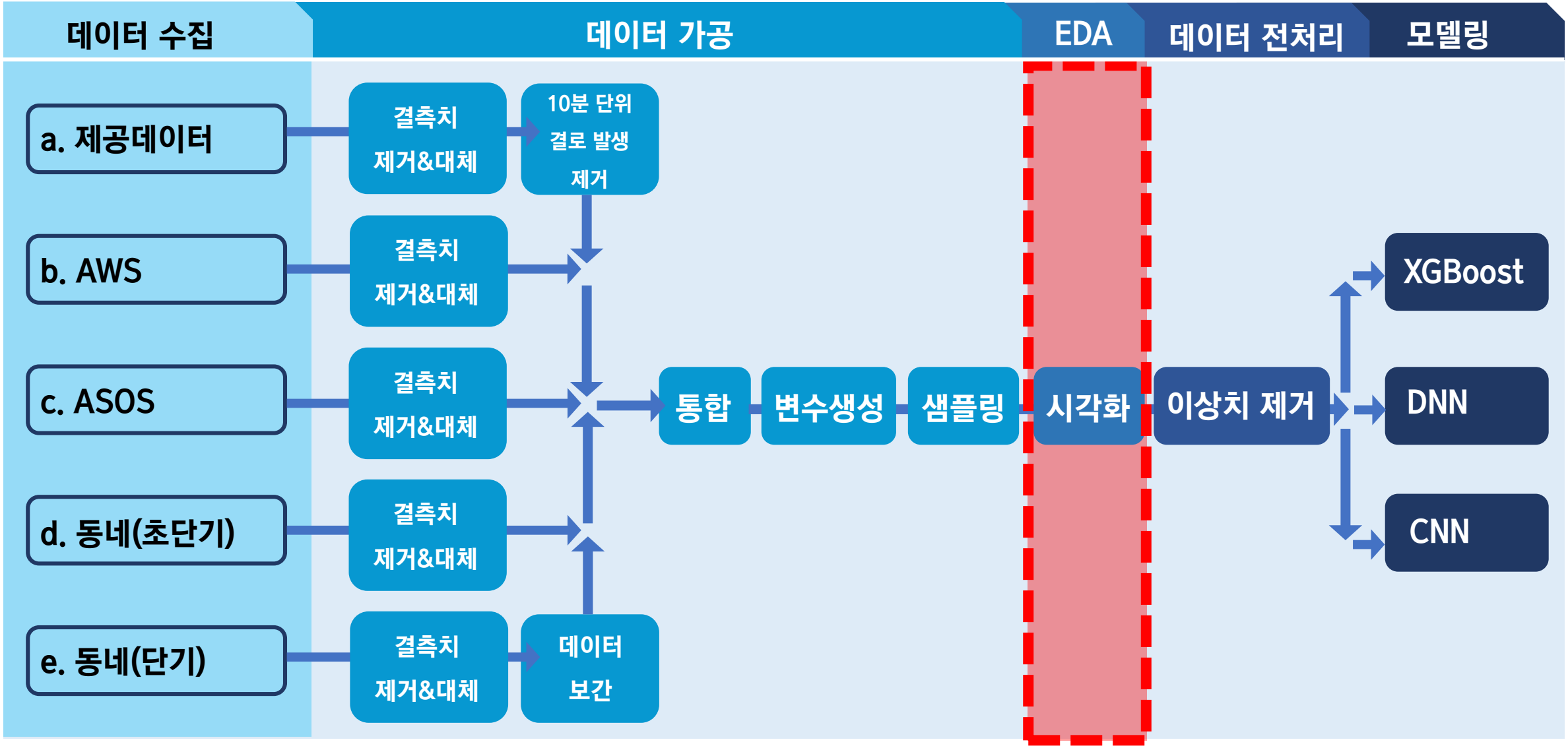


6. 샘플링(Under Sampling)

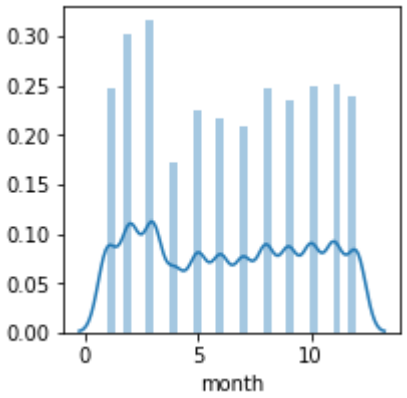
year	month	hour_interval	plant	loc	population	sample
2016	7	1	1	1	18	10
2016	7	1	1	2	18	10
2016	7	1	1	3	18	10
2016	7	1	2	1	18	10
…	…	…	…	…	…	…
2019	3	6	1	2	81	43
2019	3	6	1	3	81	43
2019	3	6	2	1	81	43
2019	3	6	2	2	81	43

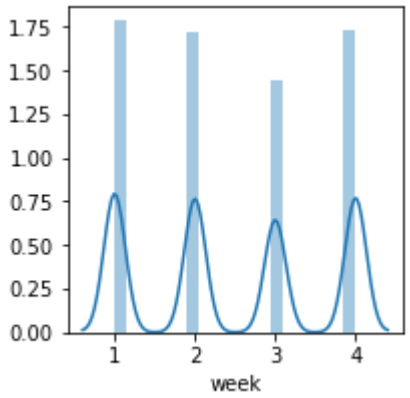
- ✓ **년도, 월, 시간대, 공장, 위치 기준**
- ✓ 결로 발생 여부가 0인 데이터를 대상으로 **6만개 데이터 층화추출**
- ✓ **불균형한 클래스 비율 조정**과 **학습시간 단축의 장점**
- ✓ **총 62353개 데이터**를 학습에 사용
- ✓ 결로 발생(positive class) 데이터는 **1216개, 1207개**
- ✓ negative / positive **클래스 비율은 약 50 : 1**

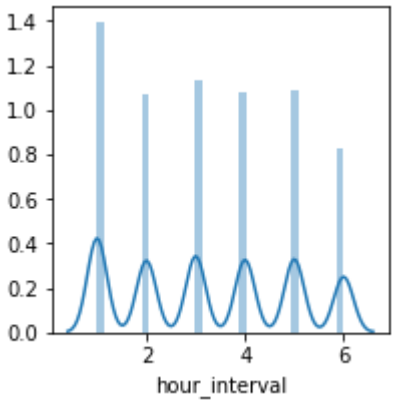
04. 탐색적 자료 분석

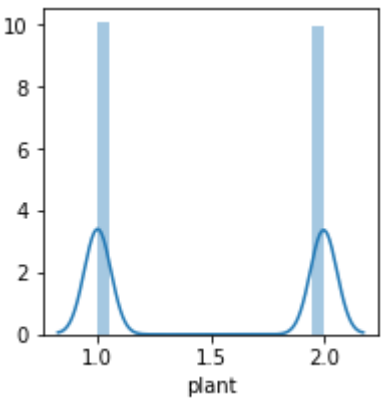

데이터 수집
데이터 가공
EDA
데이터 전처리
모델링
a. 제공데이터
b. AWS
c. ASOS
d. 동네(초단기)
e. 동네(단기)
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
10분 단위 결로 발생 제거
데이터 보간
통합
변수생성
샘플링
시각화
이상치 제거
XGBoost
DNN
CNN

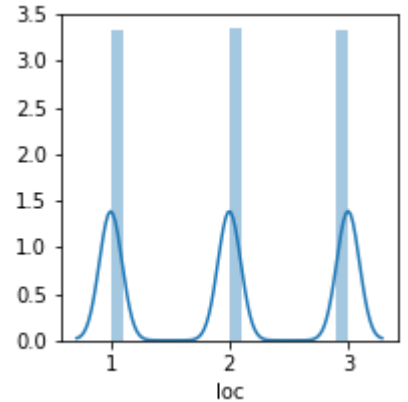
1. 주요 변수 분포(현대제철)

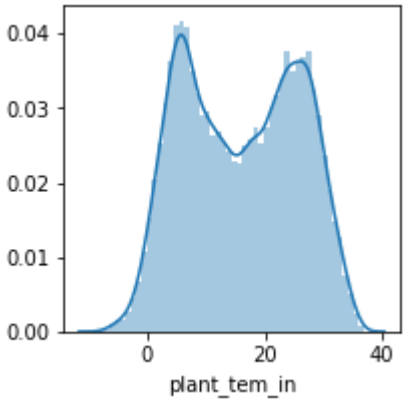


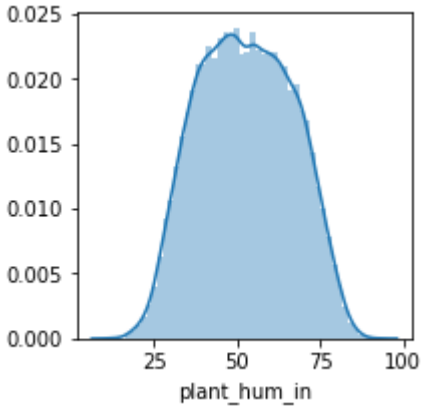


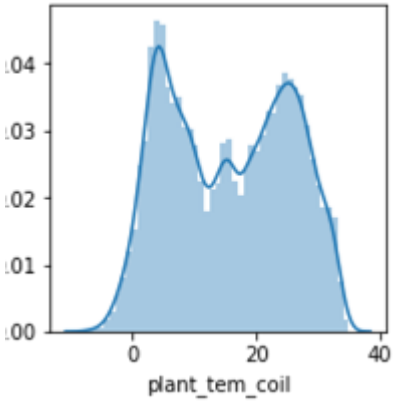


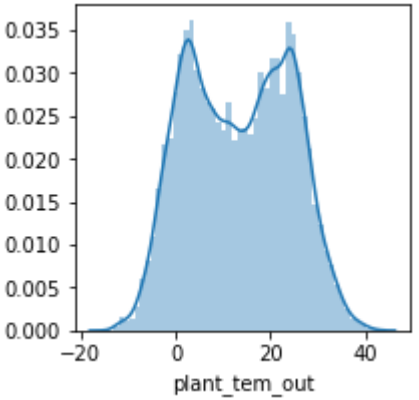


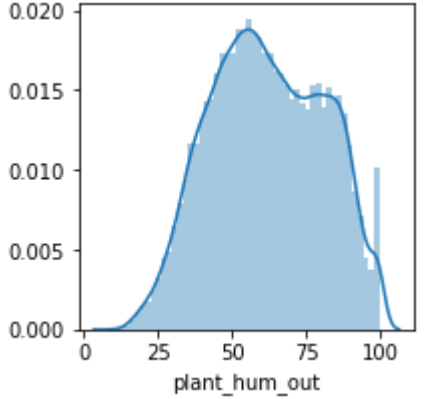


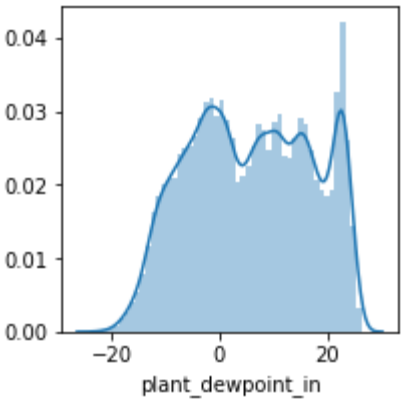


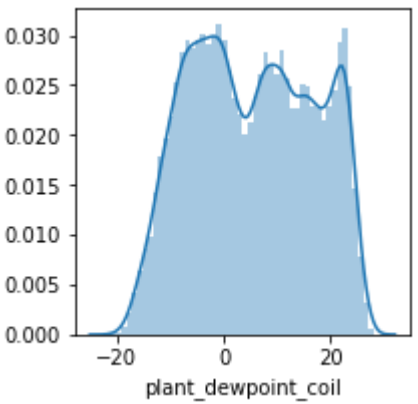


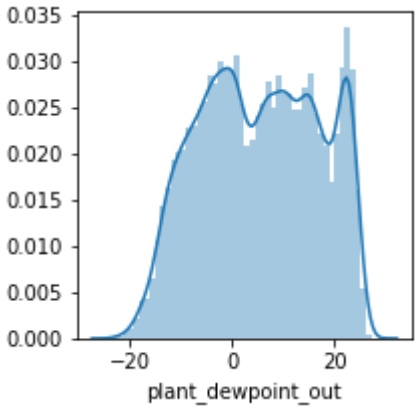








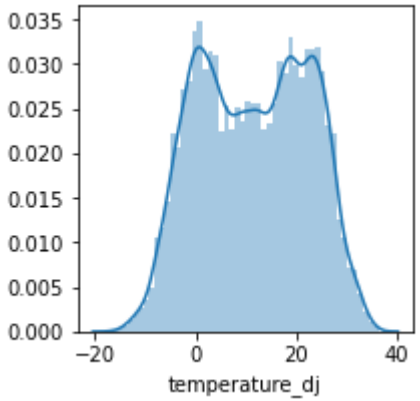


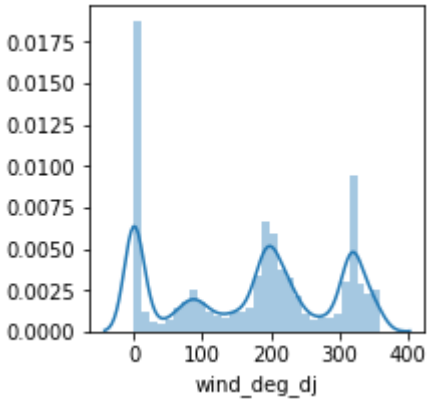


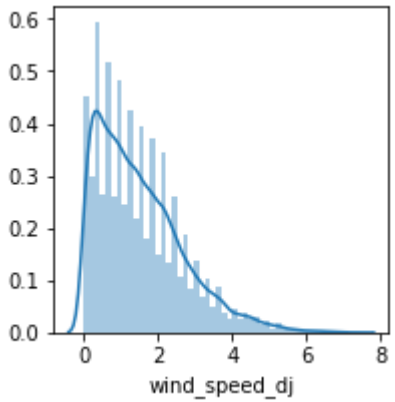
현대제철 데이터에

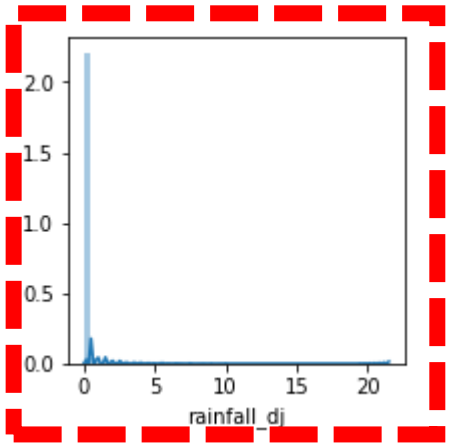
특별한 이상치는 존재하지 않음.

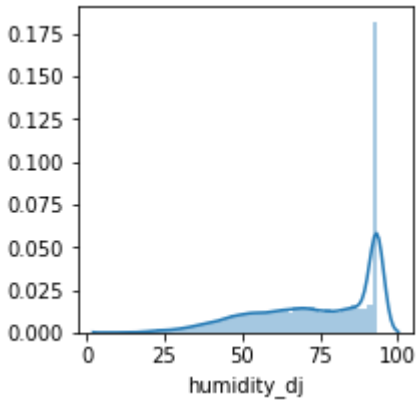
2. 주요 변수 분포(AWS)

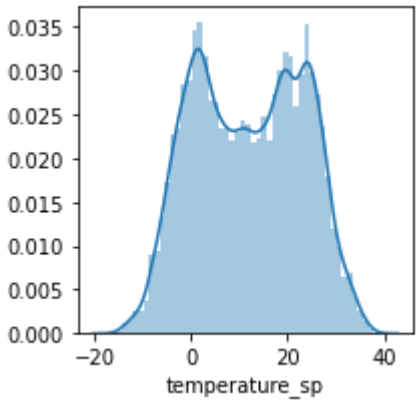


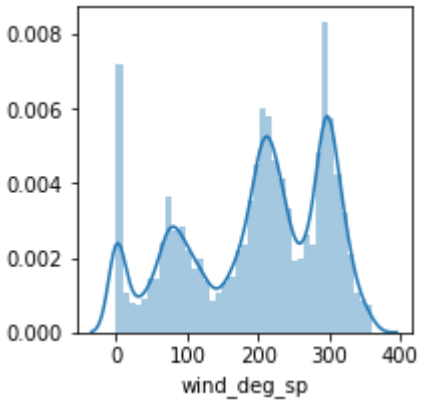


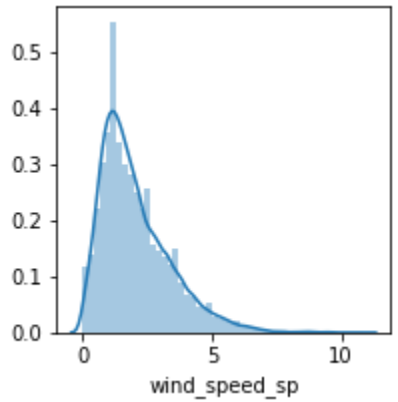


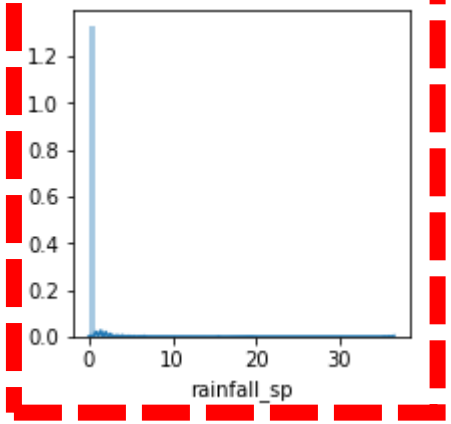


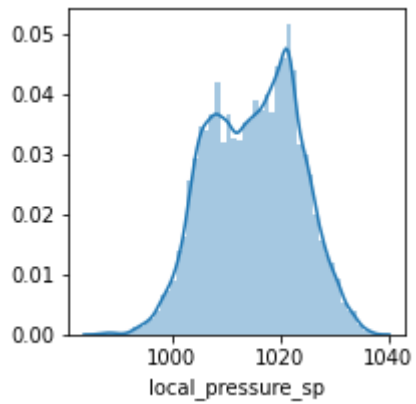


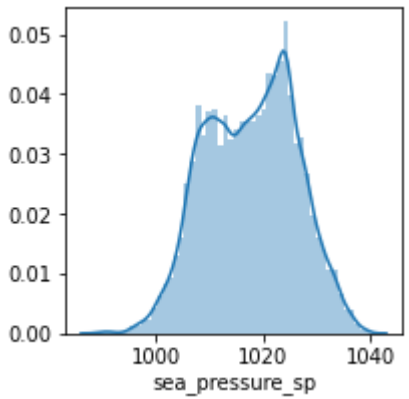








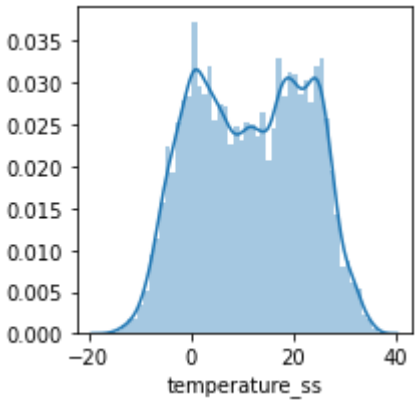


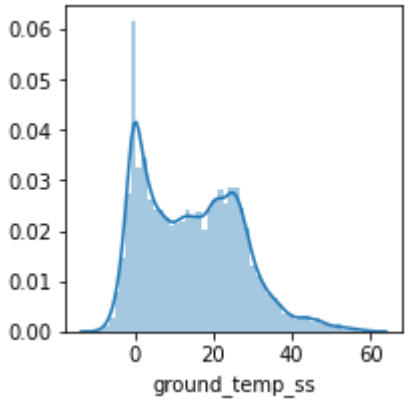


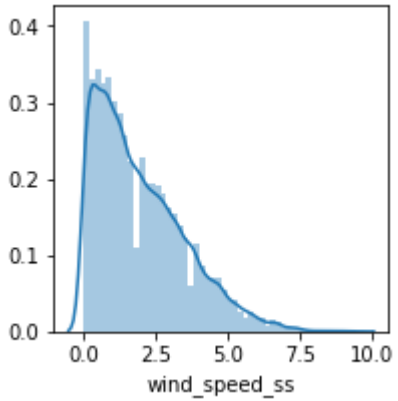
AWS의 당진, 신평

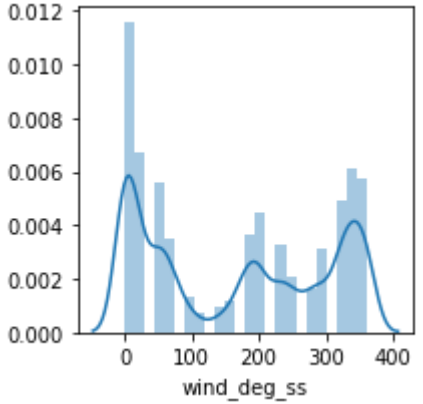
강수량 변수의 왜도가 높음.

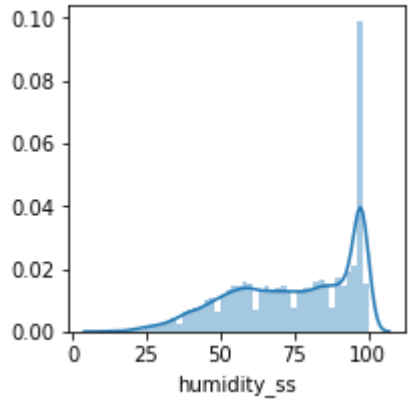
3. 주요 변수 분포(ASOS)

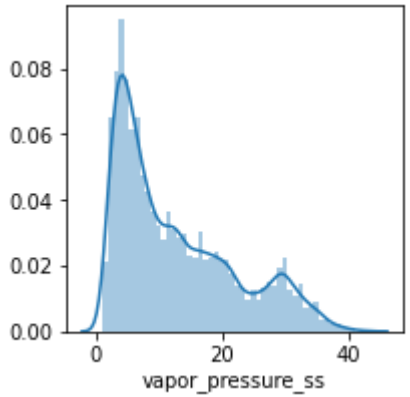


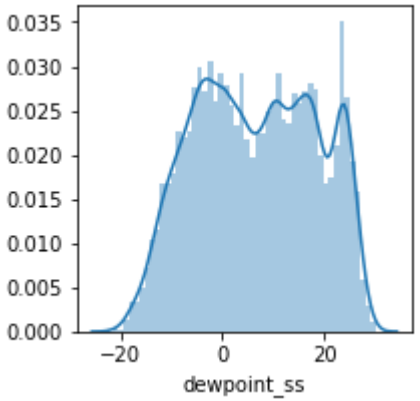


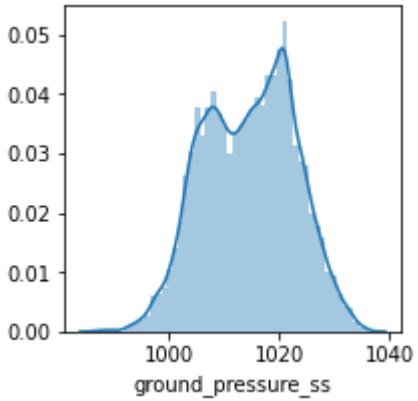


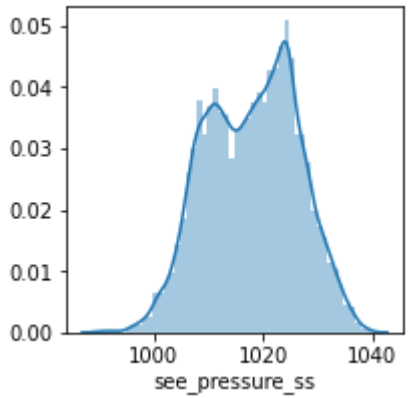










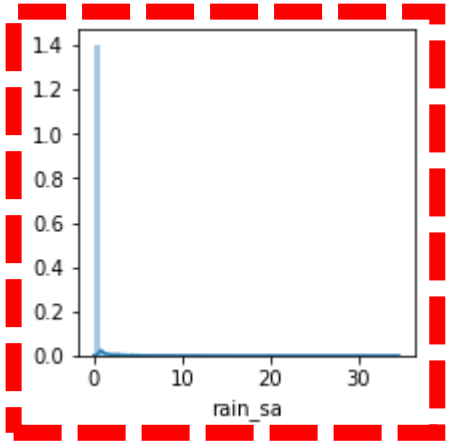


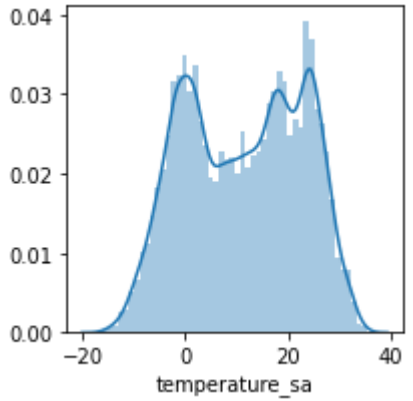
ASOS 데이터에

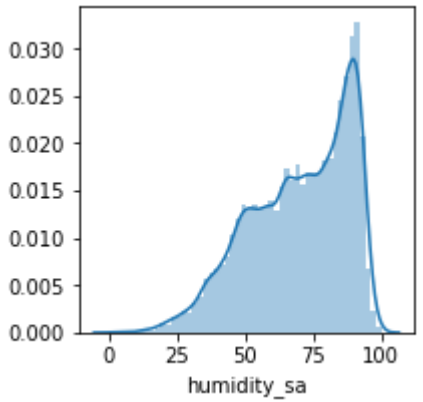
특별한 이상치는

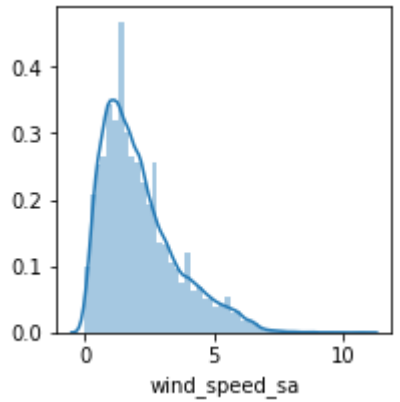
존재하지 않음.

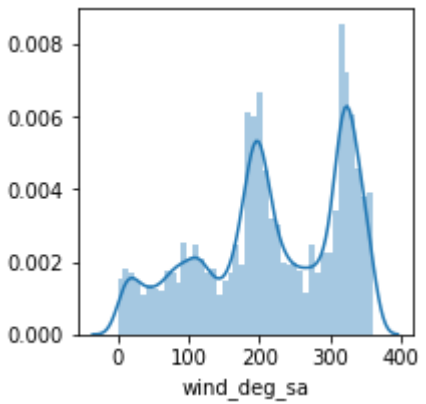
4. 주요 변수 분포(초단기예보)

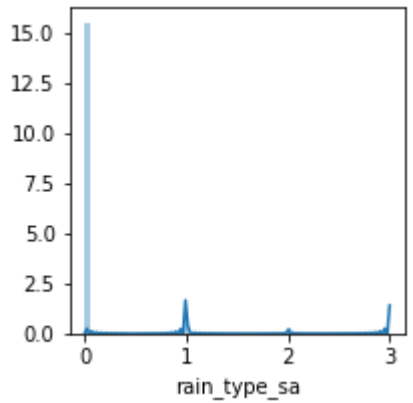


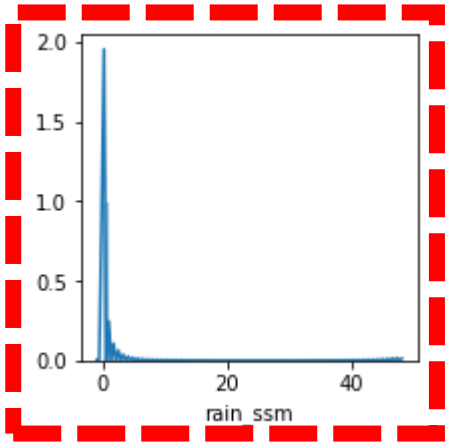


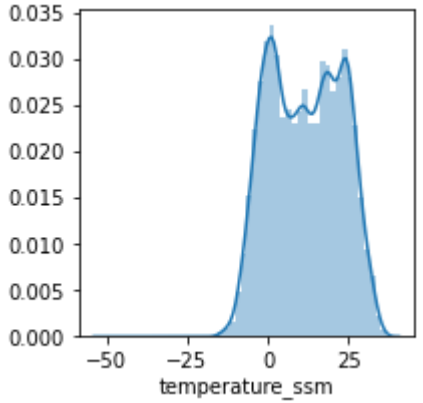


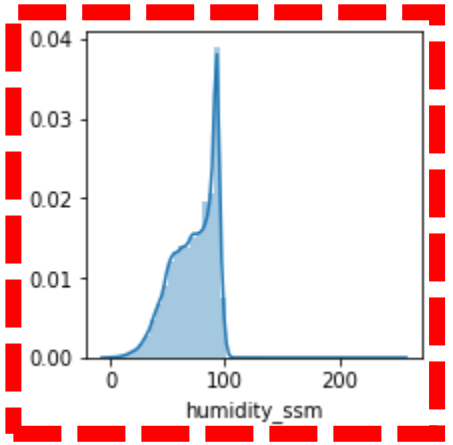


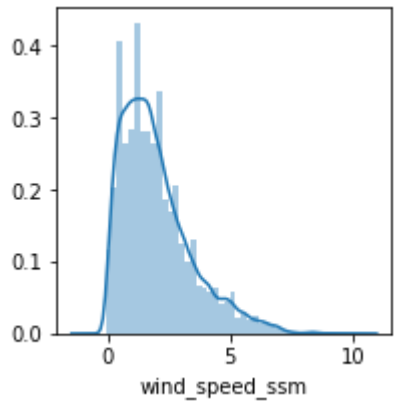


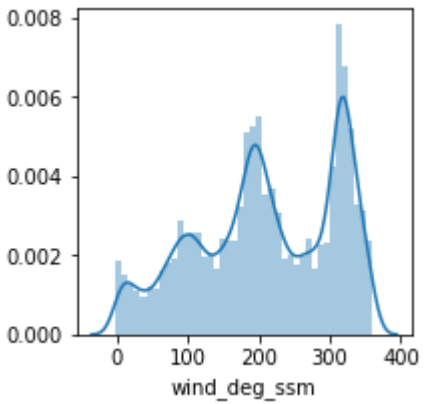


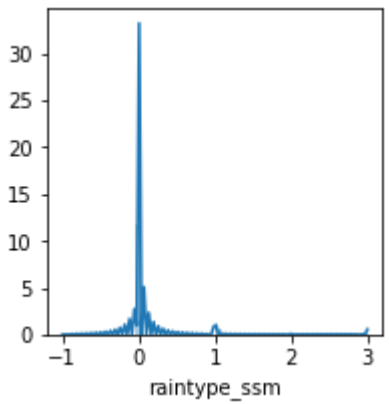








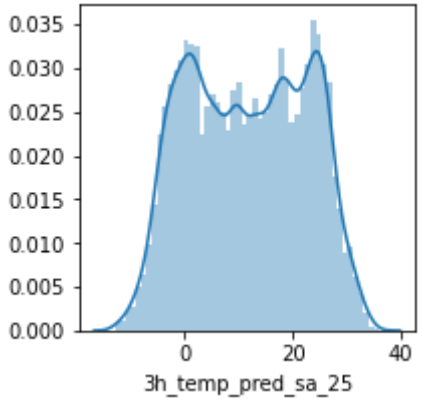


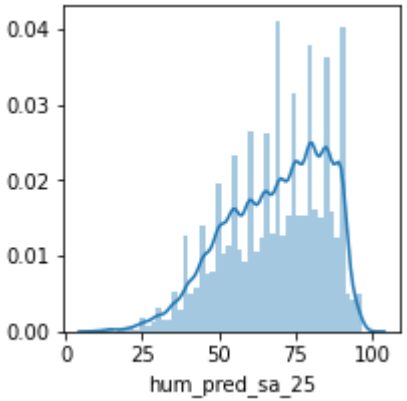


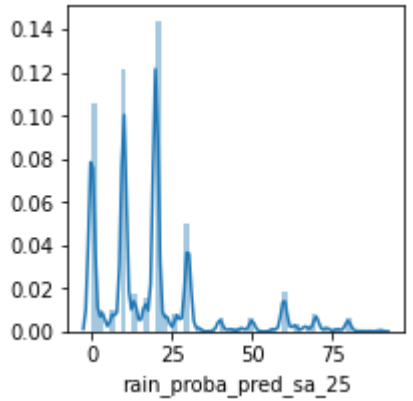
초단기예보의

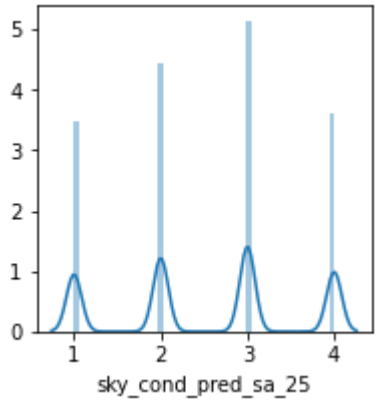
강수량, 습도 변수의 왜도가 높음.

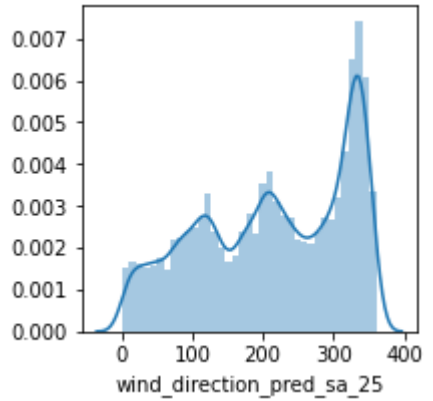
5. 주요 변수 분포(단기예보)

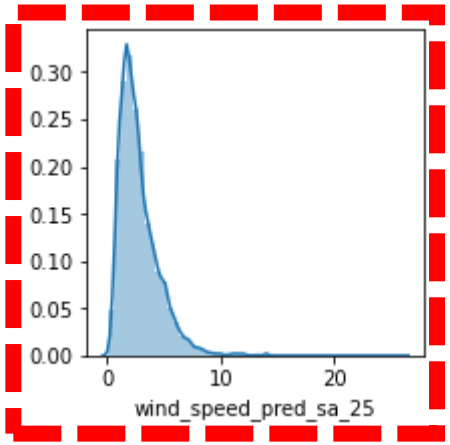


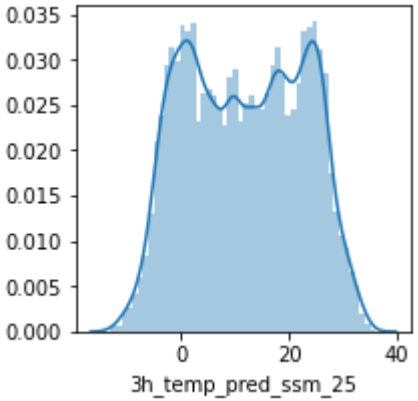


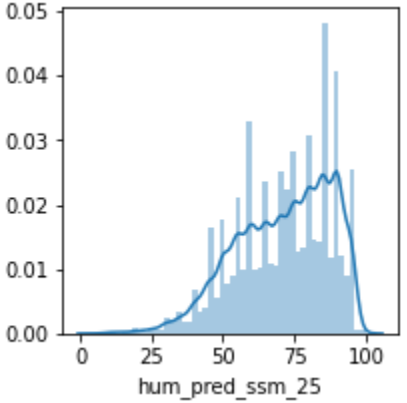


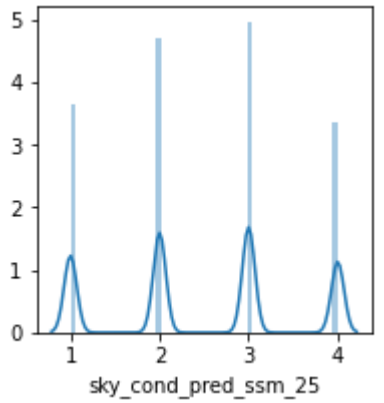


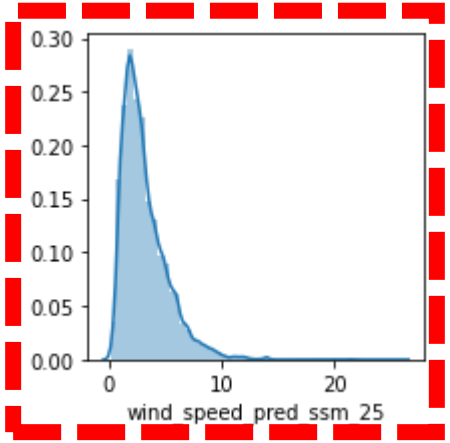


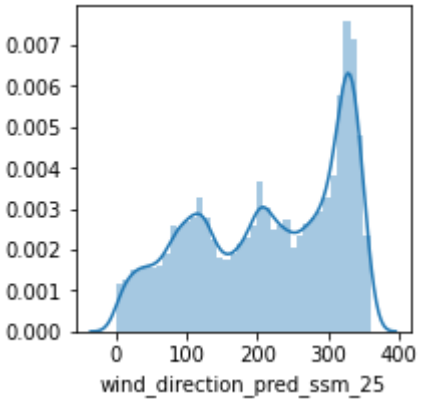






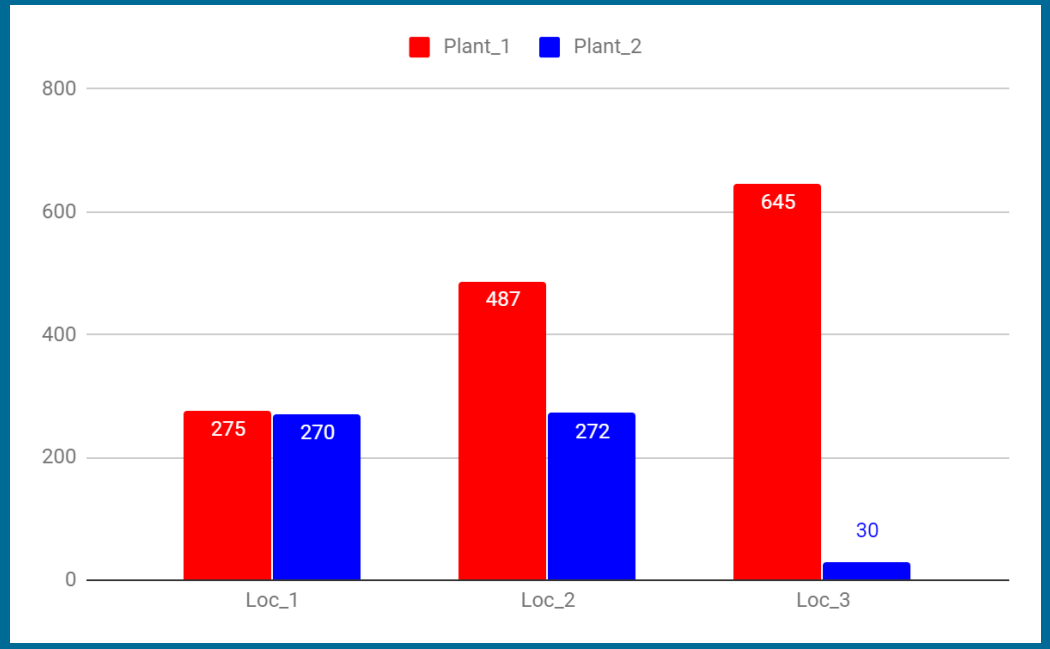


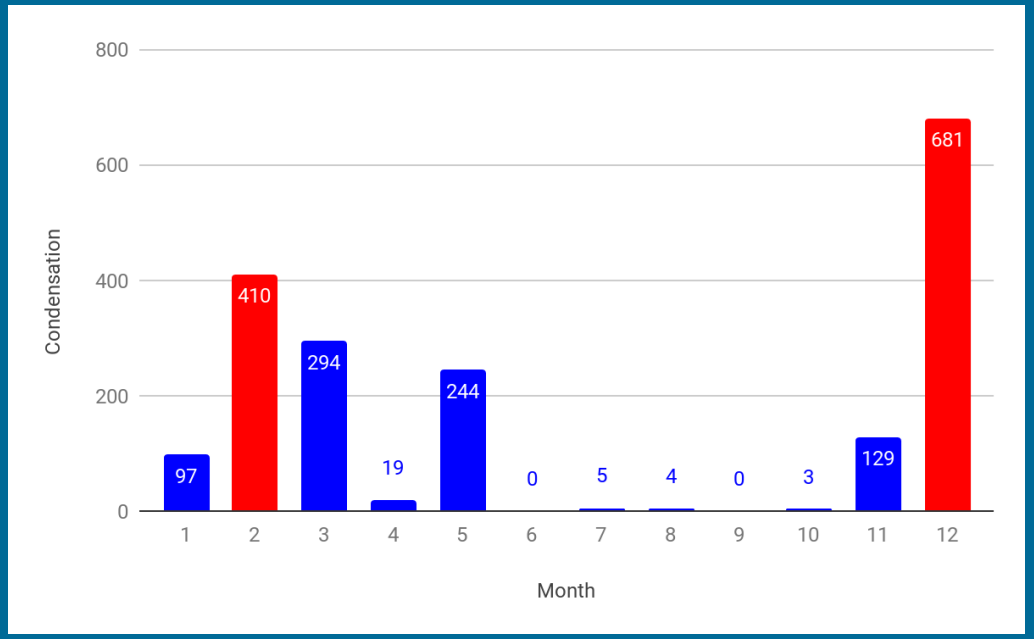




단기예보의 풍속 변수의 왜도가 높음.

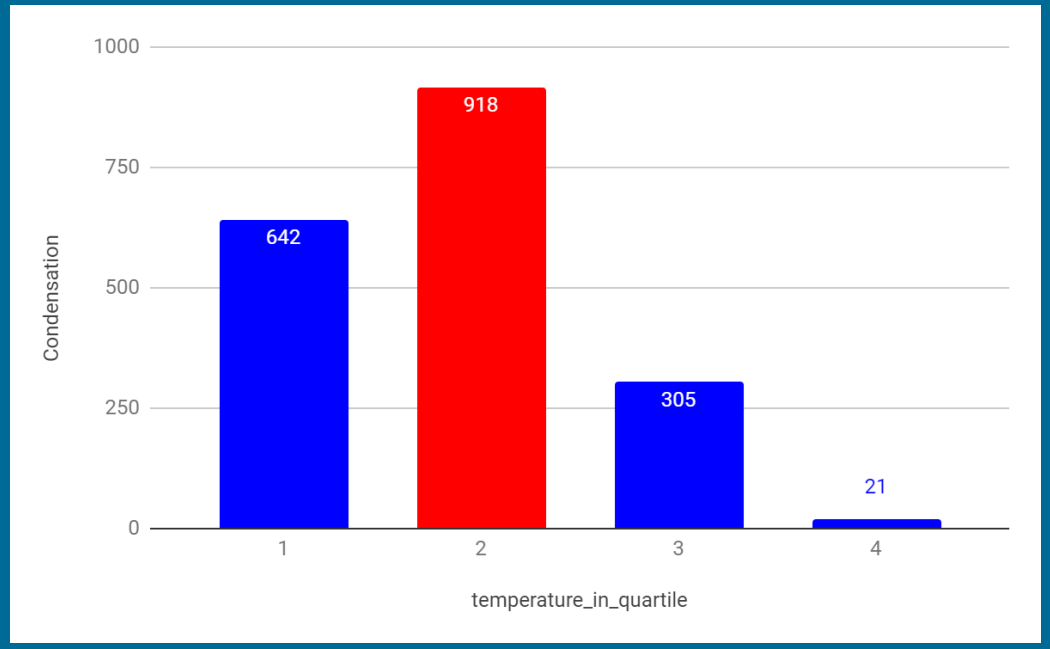
6. 공장(Plant), 위치(Loc), 월(Month)별 결로발생

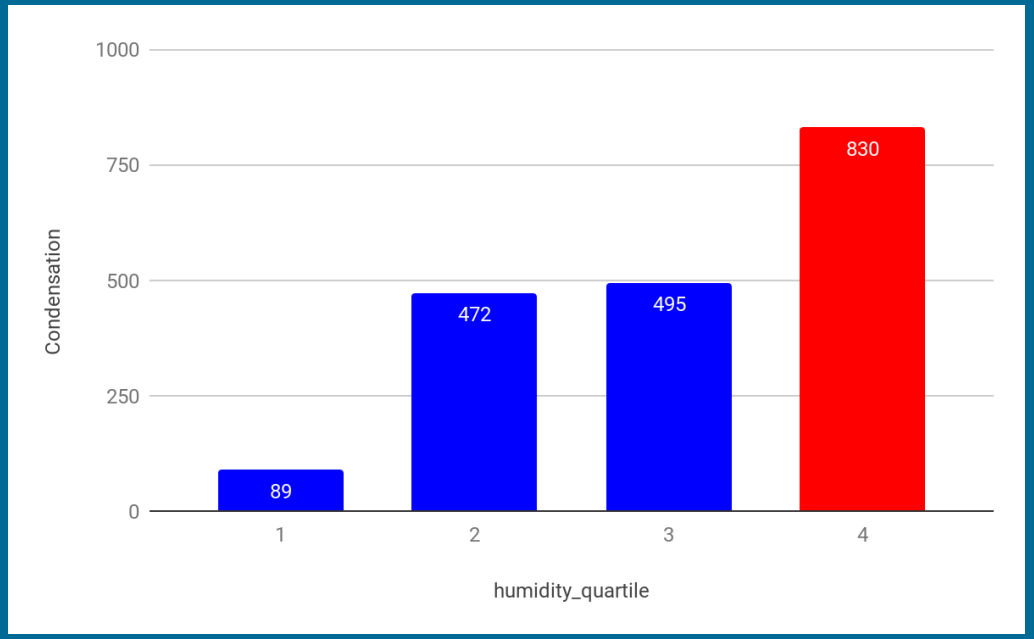




- ✓ Plant_1의 결로 발생이 상대적으로 높음.
- ✓ 특히 Loc_2, Loc_3 위치에서 두드러짐.
- ✓ Plant_2의 Loc_3 위치의 결로가 상당히 적음.
- ✓ 겨울인 12월과 2월에 결로가 상대적으로 많음.
- ✓ 일교차가 큰 봄 계절인 3월과 5월의 결로수도 눈여겨 볼 필요가 있음.

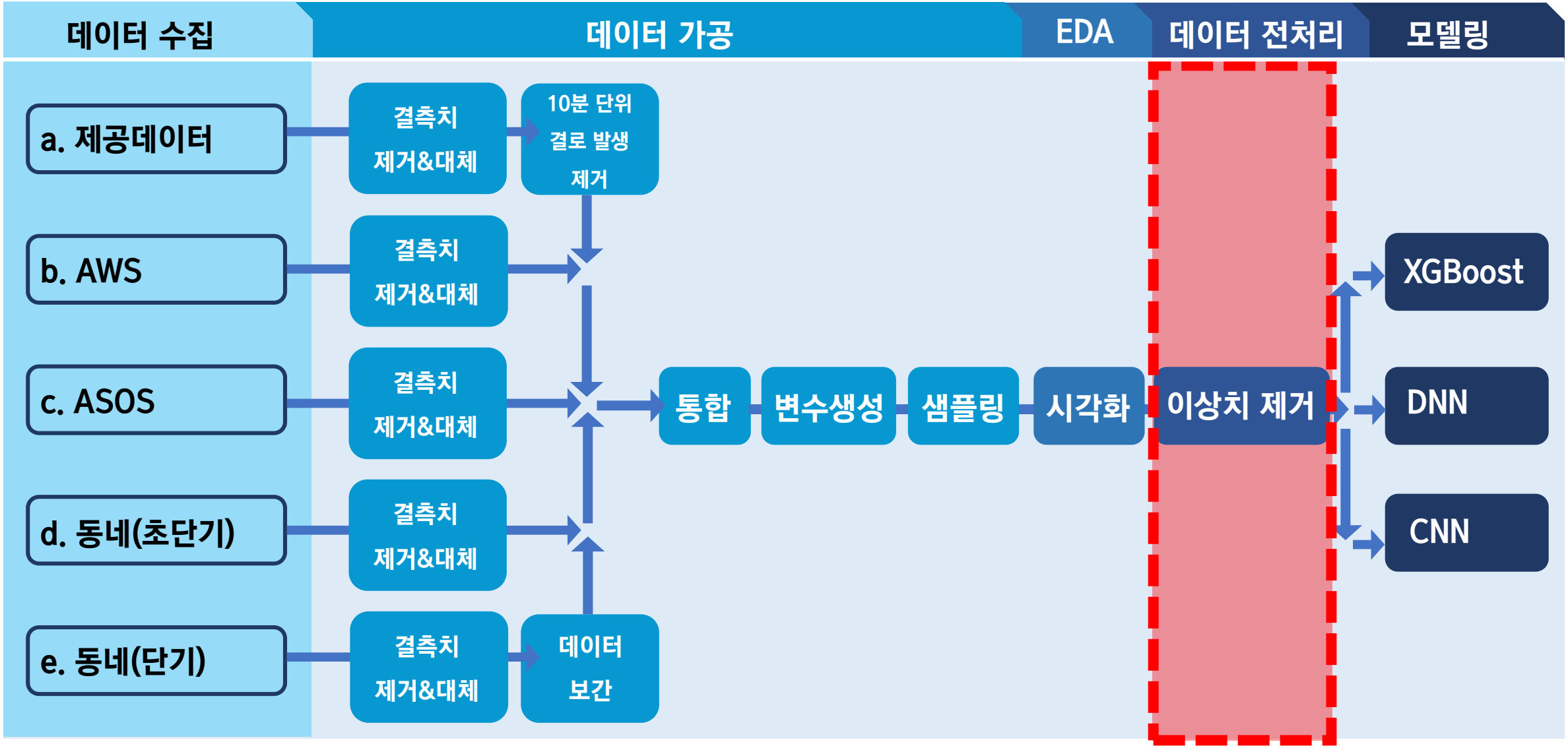
7. 공장 내부 기온(Temperature), 습도(Humidity) 분위수별 결로발생





- ✓ 공장 내부 기온 분위수가 낮을 수록 결로가 상대적으로 많이 발생함.
- ✓ 특히, 공장 내부 기온 2분위수에서 가장 높음.
- ✓ 공장 내부 습도 분위수가 높을 수록 결로가 많이 발생하는 경향이 있음.
- ✓ 공장 내부 습도 분위수가 4일 경우 결로가 가장 많이 발생함.

8. 이상치 제거


데이터 수집
데이터 가공
EDA
데이터 전처리
모델링
a. 제공데이터
b. AWS
c. ASOS
d. 동네(초단기)
e. 동네(단기)
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
10분 단위 결로 발생 제거
데이터 보간
통합
변수생성
샘플링
시각화
이상치 제거
XGBoost
DNN
CNN

8. 이상치 제거

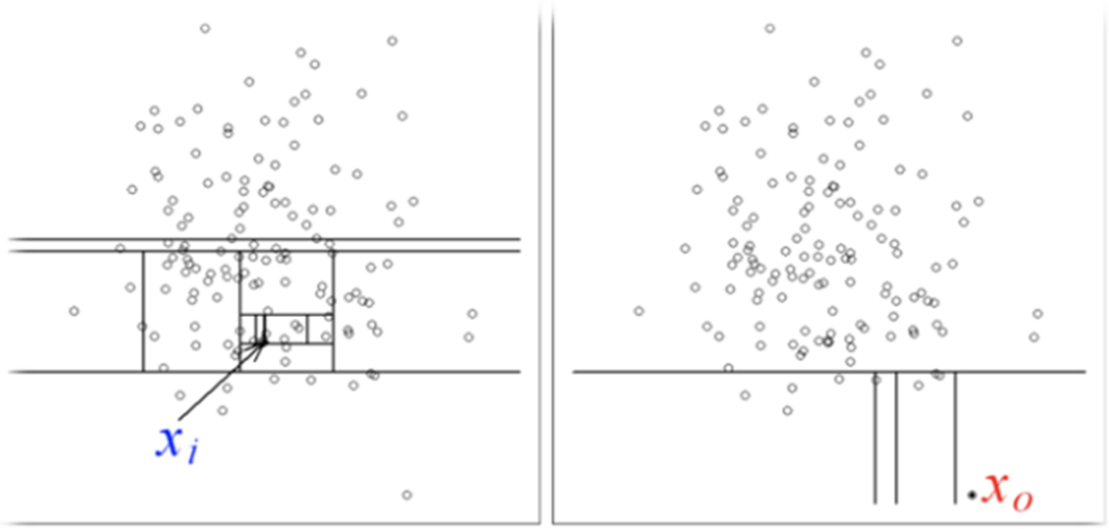
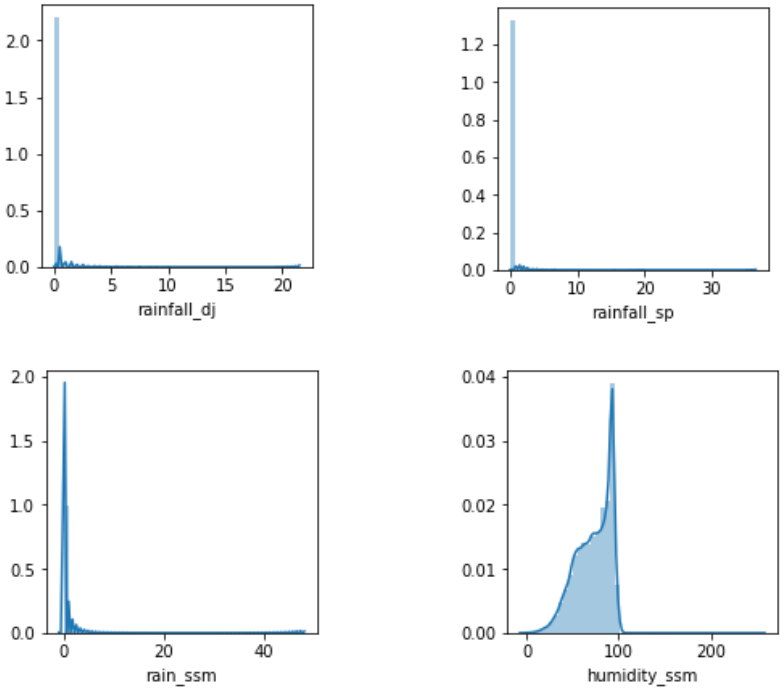


Figure 1 Identifying normal vs. abnormal observations



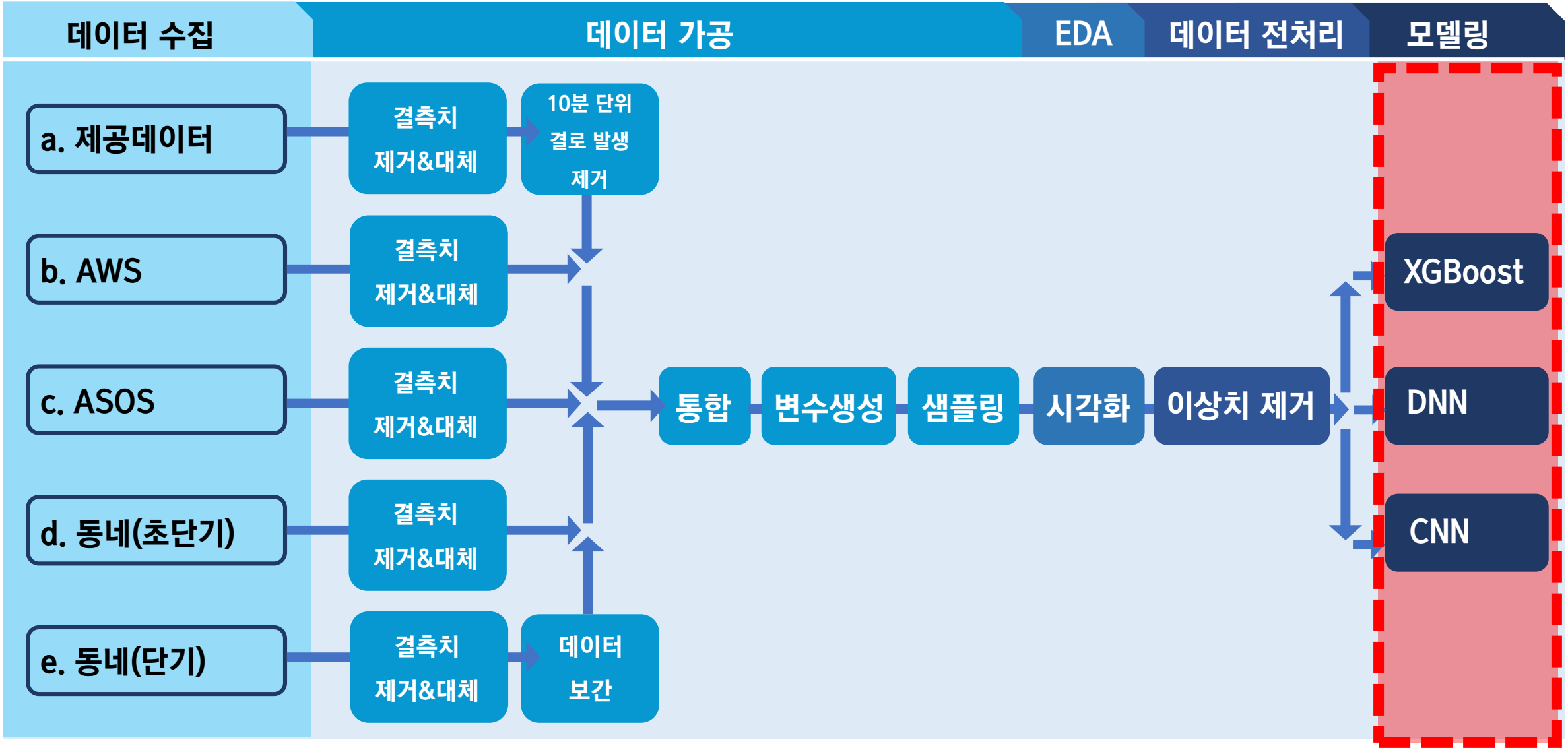
✓ Isolation Forest는 **정상 데이터**의 경우,
비정상 데이터에 비해 **더 많은 이진 분할이 필요하다는 개념에 착안**하여
Tree 모델로 이상치를 탐지함.

✓ 위 그림을 보면 왼쪽 그림(정상 데이터) 보다
오른쪽 그림(비정상 데이터)에서 **더 적은 분할 수로 나눠진 영역에**
이상치가 있음을 볼 수 있음.

✓ EDA 과정에서 강수량, 습도 등의 데이터의 분포가 치우쳐
왜도가 높은 변수들이 존재함을 볼 수 있음.

✓ **왜도 2.0을 넘는 변수를 대상**으로 Isolation Forest를 활용해
1%의 이상치를 제거함.

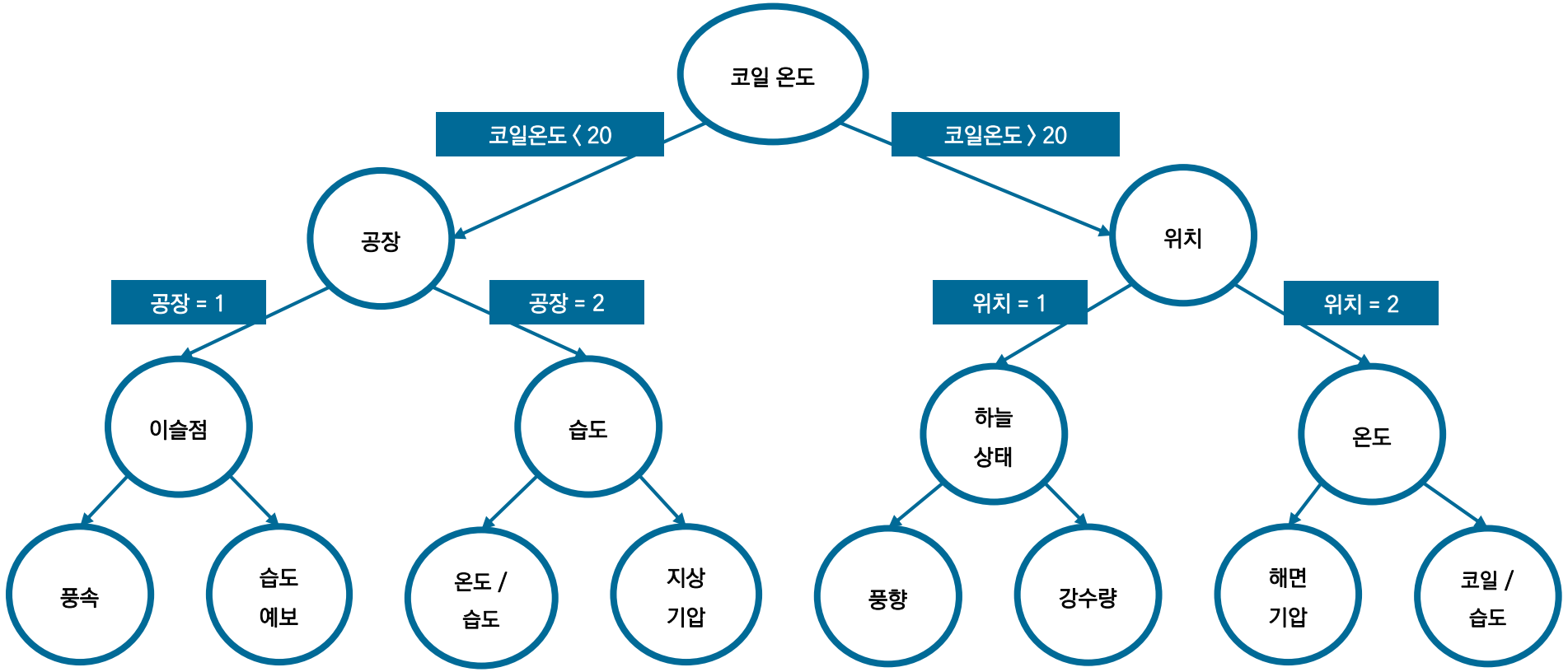
✓ **총 624개 row가 이상치로 판정**되 제거됨.


데이터 수집
데이터 가공
EDA
데이터 전처리
모델링
a. 제공데이터
b. AWS
c. ASOS
d. 동네(초단기)
e. 동네(단기)
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
결측치 제거&대체
10분 단위 결로 발생 제거
데이터 보간
통합
변수생성
샘플링
시각화
이상치 제거
XGBoost
DNN
CNN

1. 모델 – XGBoost

✓ XGBoost (eXtreme Gradient Boosting)

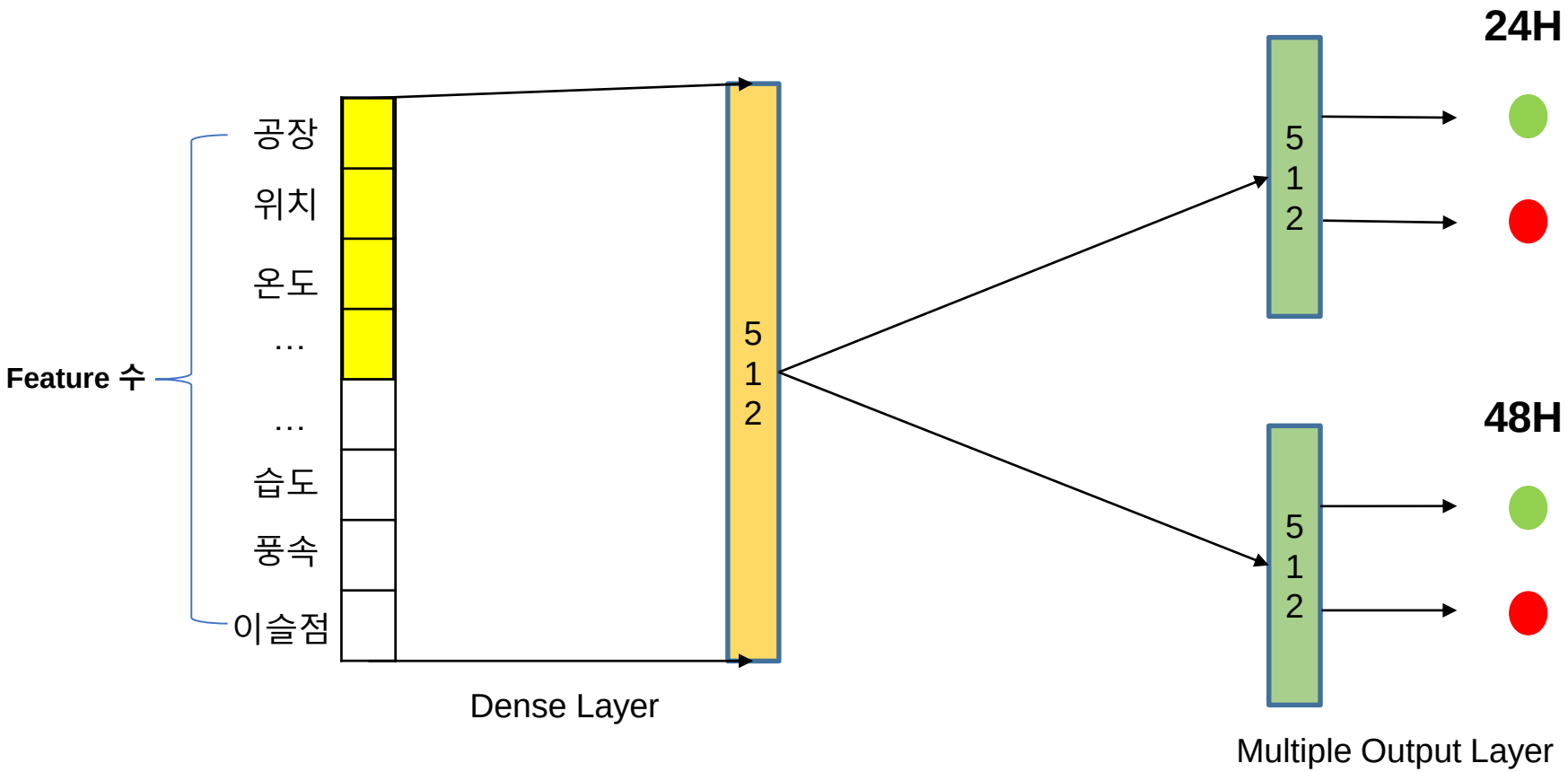
- Gradient Boosting 계열의 머신러닝 알고리즘.
- 학습 속도가 느리고 과적합 이슈가 있는 Gradient boosting 알고리즘의 단점을 보완한 모델
- 정형 데이터의 분류와 회귀 영역에서 뛰어난 예측 성능을 발휘함.
- Stratified 5 fold, Bayesian Optimization을 활용하여 하이퍼파라미터 튜닝(Train set : Test set = 80 : 20)



2. 모델 – DNN

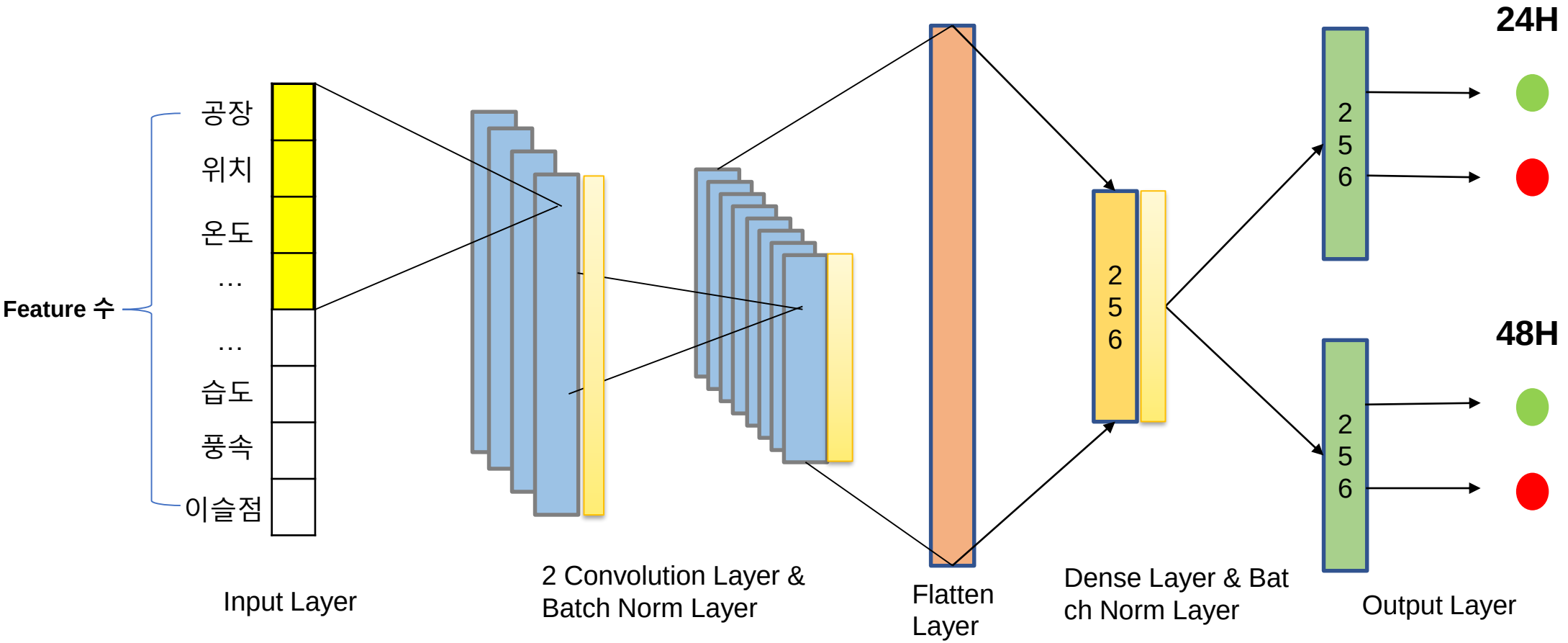
✓ DNN (Deep Neural Network)

- 입력층과 출력층 사이에 여러 개의 은닉층으로 이루어진 인공신경망
- 입력 변수들 간의 비선형 조합이 가능하다는 장점이 있음.
- 현재 다양한 분야에서 머신러닝 기법에 비해 우수한 경우가 많음.



✓ CNN (Convolutional Neural Network)

- 데이터로부터 특징을 추출하는 Convolution Layer를 포함하는 Neural Network
- CNN은 이미지의 특성인 픽셀 간 상관관계가 높은 특성을 잘 반영함.
- 본 데이터 역시 상관관계가 매우 높은 Feature들로 구성되어 있으므로 CNN 모델을 사용함.



4. 모델별 성능 비교

Validation Score

Submission Score

Model	Accuracy	Precision	Recall	F1_Score	ROC_AUC	CSI	Average_AUC
XGBoost_24	0.9871	0.6036	0.9833	0.7480	0.9853	70.58	0.9906
XGBoost_48	0.9885	0.6320	0.9834	0.7695	0.9860	70.58	0.9906
DNN_24	0.9936	0.7066	0.9553	0.8124	0.9747	45.28	0.9576
DNN_48	0.9935	0.6932	0.9831	0.8130	0.9883	45.28	0.9576
CNN_24	0.9827	0.4557	0.9772	0.6206	0.9963	55.50	0.9781
CNN_48	0.9835	0.4644	0.9944	0.6331	0.9970	55.50	0.9781

최종 모델의 CSI(모델 종합 정확도)는 70.58%, 평균 AUC는 0.9906, Cut-off는 0.30

5. 최종 모델 해석

XGBoost_24 Permutation Importance TOP 9

Weight	Standard Error	Feature Name
0.0073	± 0.0026	Pred_ssm_13_dewpoint
0.0049	± 0.0022	Plant_tem_coil
0.0036	± 0.0010	Plant_tem_in__Plant_tem_coil__ratio
0.0021	± 0.0009	Plant_hum_in__Plant_hum_out__ratio
0.0013	± 0.0003	Plant_tem_coil__Plant_tem_out__ratio
0.0010	± 0.0005	Plant_tem_in__Plant_tem_out__ratio
0.0009	± 0.0008	loc
0.0009	± 0.0007	Plant_tem_coil__Plant_hum_out__ratio
0.0009	± 0.0003	Plant_tem_coil_day_max

XGBoost_48 Permutation Importance TOP 9

Weight	Standard Error	Feature Name
0.0012	± 0.0004	Plant_tem_coil
0.0010	± 0.0004	Wind_direction_pred_ssm_25
0.0009	± 0.0004	Pred_ssm_37_dewpoint
0.0007	± 0.0003	Plant
0.0007	± 0.0004	Plant_tem_coil_day_min
0.0007	± 0.0004	season
0.0005	± 0.0003	Hum_pred_ssm_37
0.0005	± 0.0002	Loc
0.0004	± 0.0002	Plant_tem_coil__Plant_hum_out__ratio

✓ Permutation Importance(순열 특성 중요도)는 학습된 모델을 Test set에 대해서 한 번에 한개의 변수를 선택해 값을 무작위로 섞어 평가지표의 변화를 측정함.

✓ 변수의 값을 무작위로 섞고 학습 시 평가지표에 변화가 없다면 선택된 변수는 학습된 모델에서 중요하지 않은 변수로 간주하는 방법론임.

✓ 최종 모델로 선택된 XGBoost 모델에 AUC를 기준으로 모든 변수에 대해 100번의 Permutation Importance를 수행함.

✓ 24시간 후와 48시간 후의 결로발생을 예측하는데 코일온도, 이슬점, 풍향, 위치, 공장 내부 기온과 습도의 비율, 코일온도와 외부 습도의 비율 등의 변수가 중요하게 작용함.

모니터링이 가능하도록 결로 발생 지수를 설계

결로 발생 지수 = 결로 예측 값(*50) + 결로 발생 환경(*50)

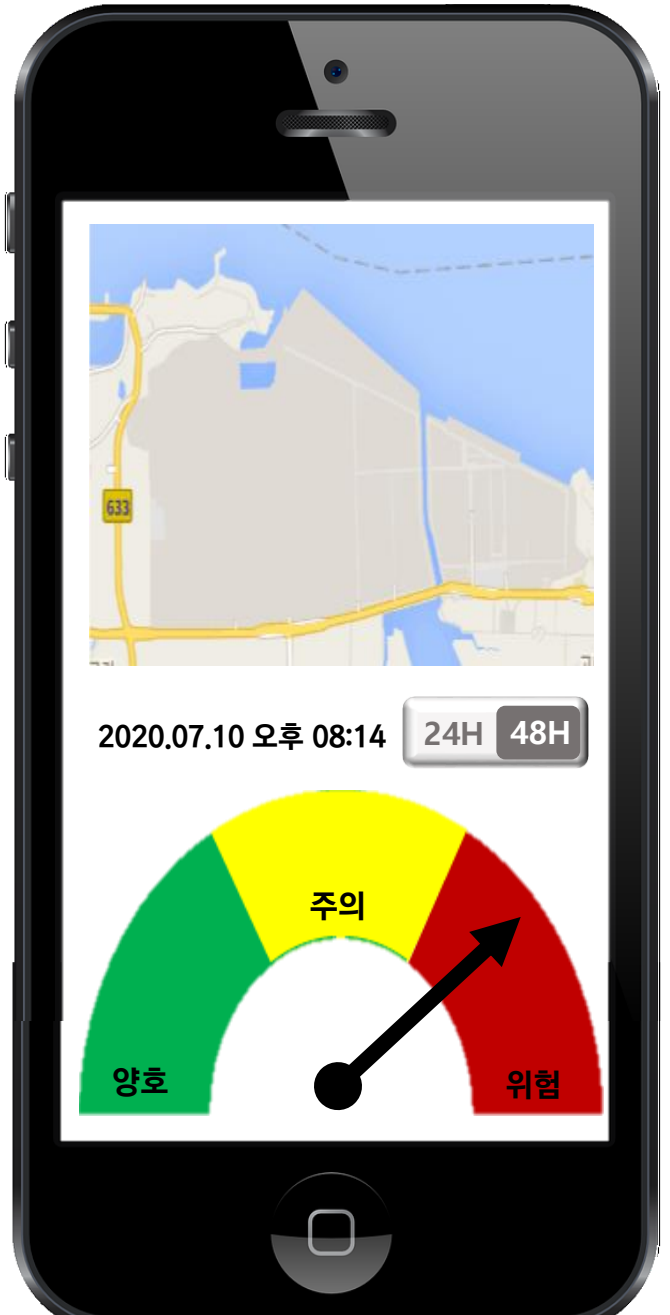
- ✓ 결로 예측 값 (%) -〉 0~1 사이의 확률값 (결로이면 1, 결로가 아니면 0에 가까운 값)
- ✓ 결로 발생 환경-〉 실내 온도 분위수(가중값)(0.25) + 실내습도 분위수(가중값)(0.25)

예시 〉 1. 결로 확률 : 0.7

2. 실내 온도 분위수 2, (가중값 1)

3. 실내 습도 분위수 4, (가중값 1)

(0.35 × 50) + (0.5 × 50) = 85 (위험)



모니터링이 가능하도록 결로 발생 지수를 설계

결로 발생 지수 = 결로 예측 값(*50) + 결로 발생 환경(*50)

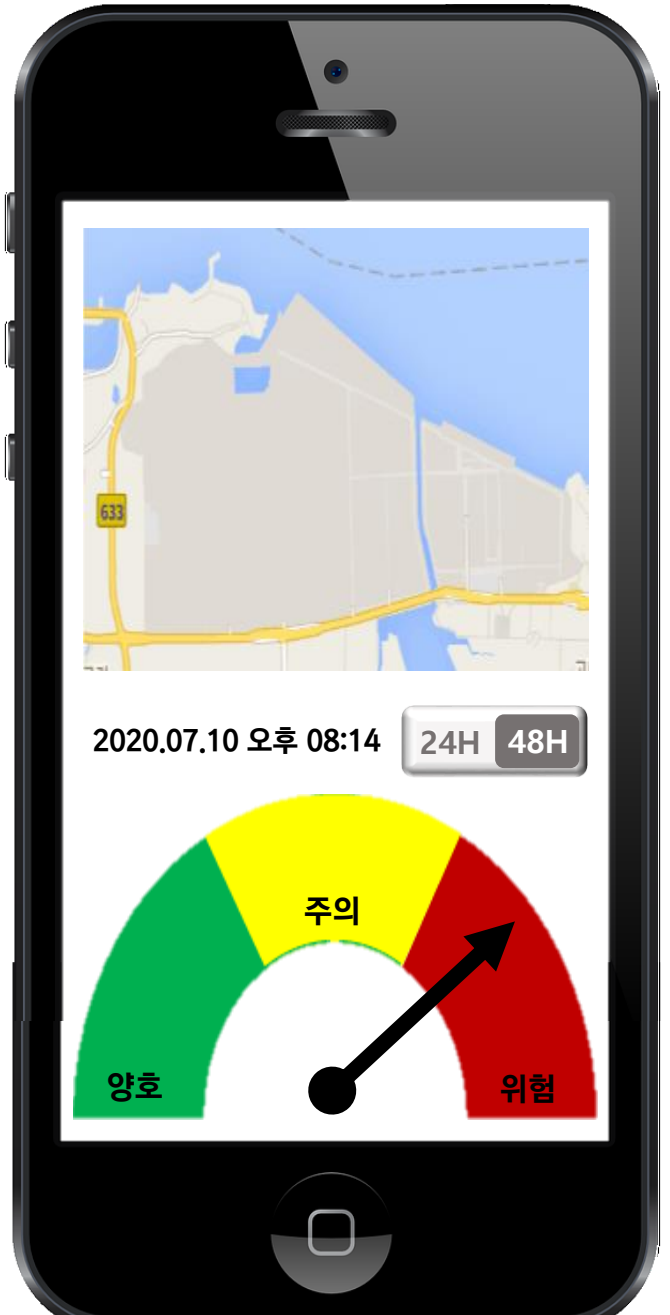
- ✓ 결로 예측 값 (%) -〉 0~1 사이의 확률값 (결로이면 1, 결로가 아니면 0에 가까운 값)
- ✓ 결로 발생 환경-〉 실내 온도 분위수(가중값)(0.25) + 실내습도 분위수(가중값)(0.25)

예시 〉 1. 결로 확률 : 0.7

2. 실내 온도 분위수 2, (가중값 1)

3. 실내 습도 분위수 4, (가중값 1)

(0.35 × 50) + (0.5 × 50) = 85 (위험)



✓ 결로 발생 위험이 높은 시점을 **사전에 모니터링**하여

결로를 방지하는 조치를 취할 수 있음.

✓ 이는 현대제철 철강제품의 **품질 손실을 방지**하고

손실 비용을 최소화하여 경영자원 운용의 효율성을 높여줄 수 있을 것으로 기대됨.

✓ 특히 최종 선정된 모델은 **겨울철 결로 발생을 예측**하는 데 좋은 성능을 보임.

참고문헌

- ✓ 주은지. (2020). 공동주택 드레스룸 결로 방지를 위한 LSTM 모델 기반 예측 제어. 서울대학교 대학원 건축학과.
- ✓ 박상현. (2009). 공동주택 창호 표면 결로 예측 방법 및 설계 기준 설정에 관한 연구. 연세대학교 대학원 건축공학과.

학습 모델 파라미터

Model	Learning rate	n_estimator	subsample	colsample_bytree	reg_lambda	reg_alpha	max_depth	max_delta_step	scale_pos_weight	early_stopping_rounds
XGBoost_24	0.2	1000	0.8	1.0	10.0	5.214	2.0	10.0	170	50
XGBoost_48	0.2	1000	0.8	1.0	6.339	1.964	2.0	10.0	170	50

Model	Batch Size
DNN	256

Model	Filter Kernel size	Stride	Dropout	Batch Size
CNN	(11, 11)	(1, 1)	0.1	256

감사합니다.

콘테스트 공모 문제해결 과정별 팀원 참여도

구분	강상규	구본아	신우석
문제이해 및 자료조사	40	30	30
데이터 전처리	50	20	30
데이터 모델링	40	20	40
분석결과 정리 및 보고서 작성	60	20	20
활용 방안 아이디어 제시	30	60	10