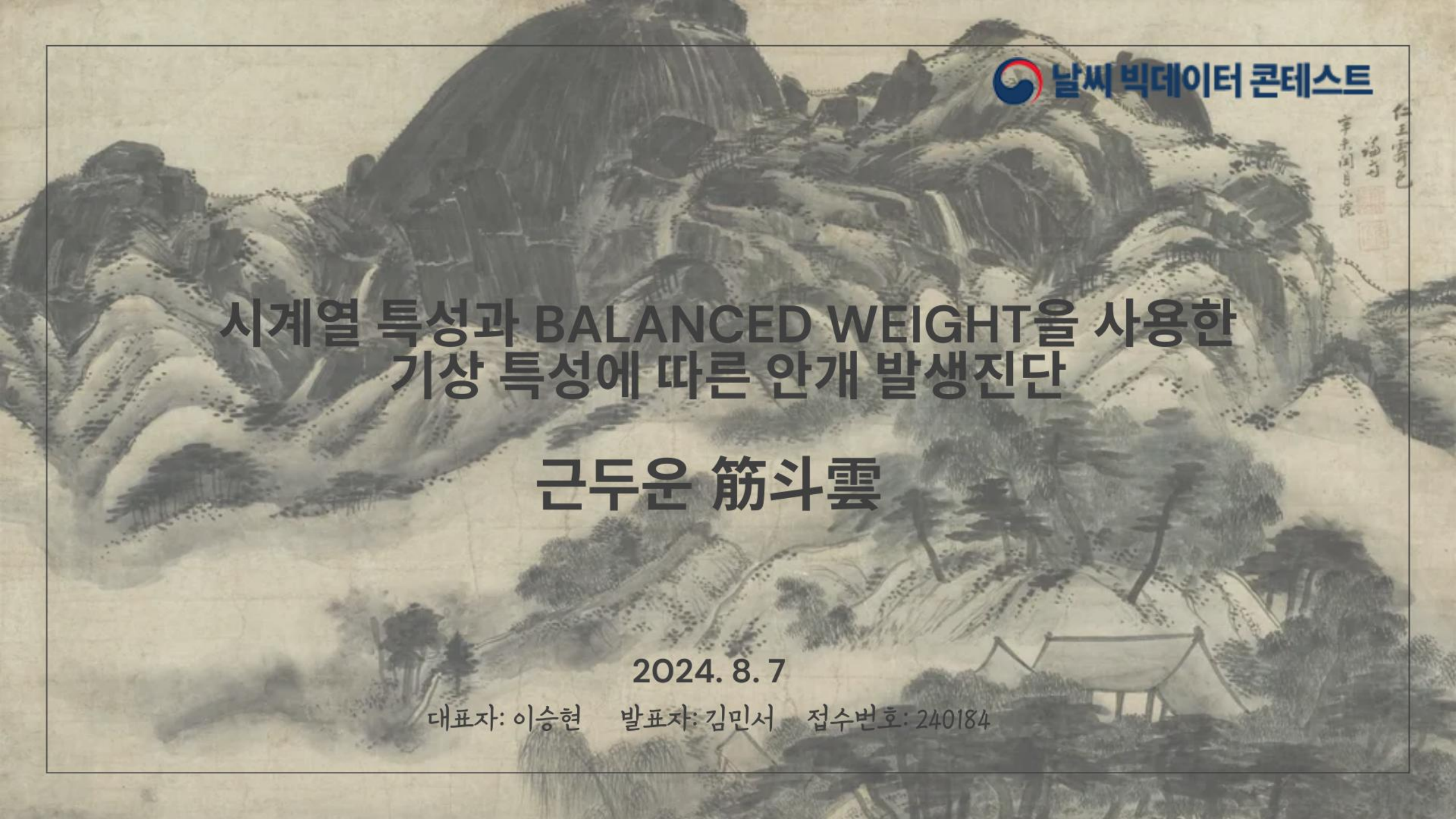


날씨 빅데이터 콘테스트

시계열 특성과 BALANCED WEIGHT을 사용한 기상 특성에 따른 안개 발생진단

근두운 筋斗雲

2024. 8. 7

대표자: 이승현 발표자: 김민서 접수번호: 240184

배경 및 목표

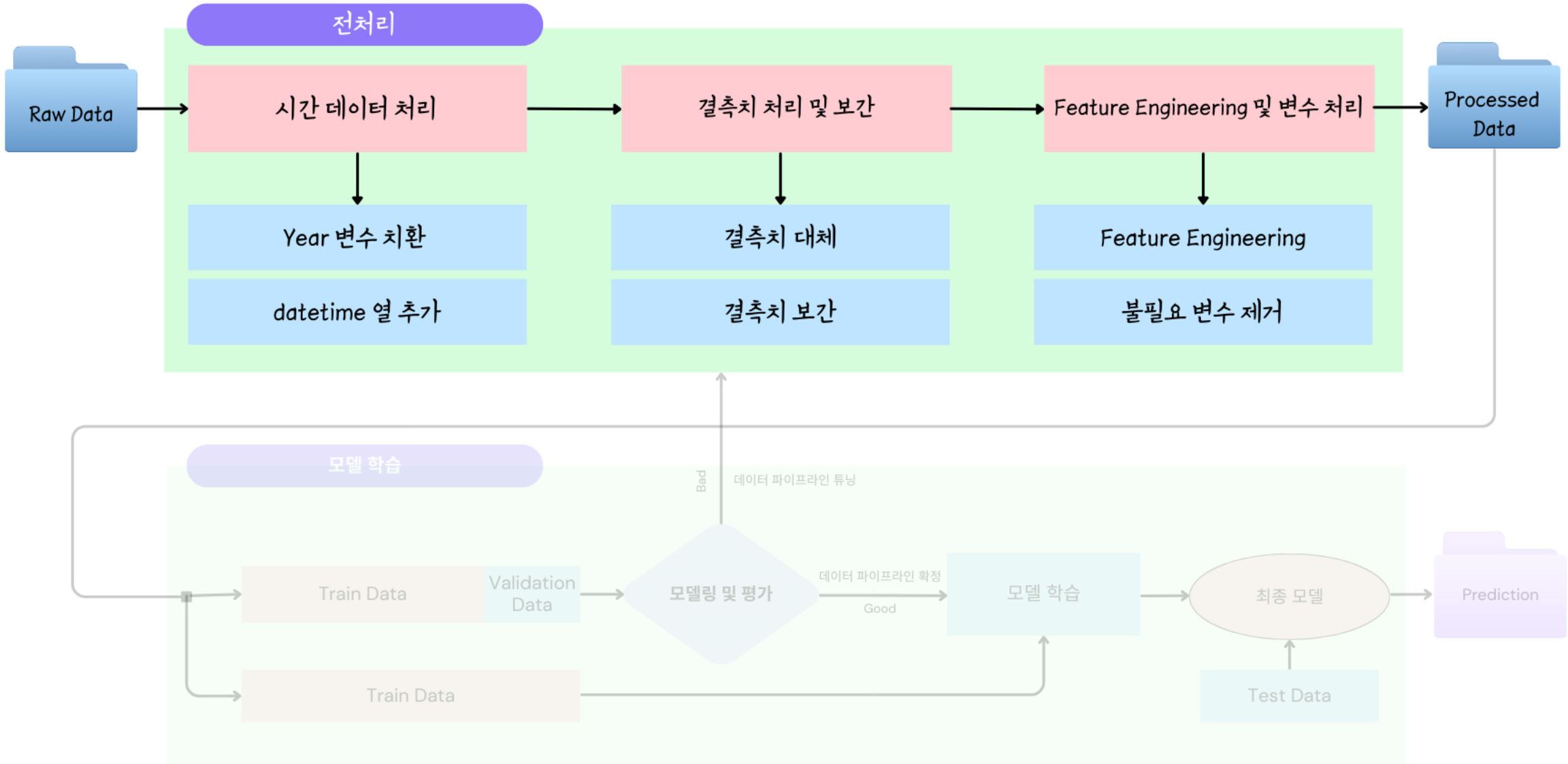
배경

안개는 항공기 사고, 교통 지체, 호흡기 문제와 같이 우리 일상 과 밀접하게 연결되어 있다.
기상관측 데이터를 분석하여 안개 발생을 예측하는 모델이 필요하다.

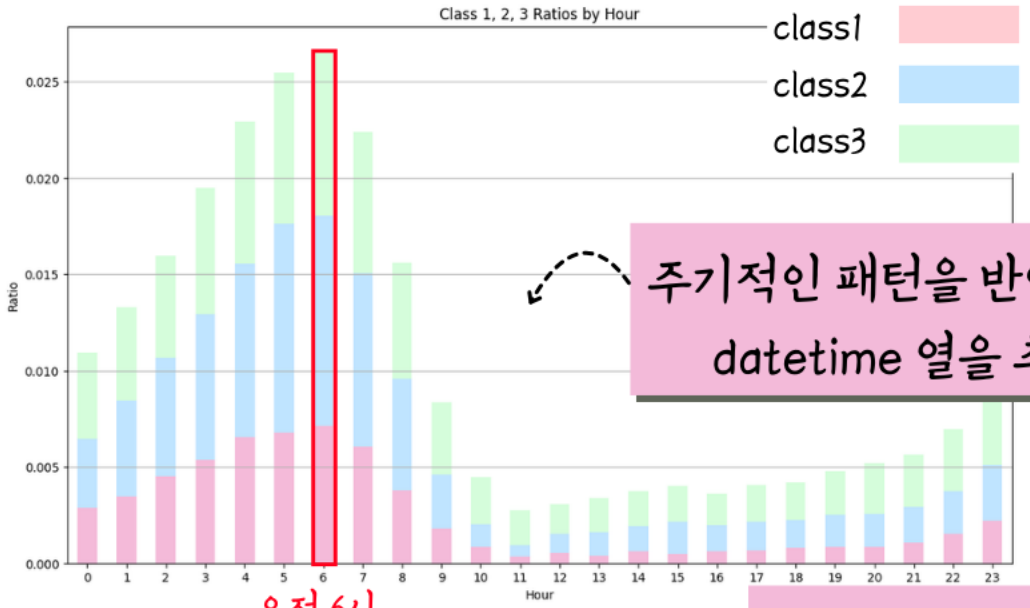
분석 목표

시간에 따른 습도, 풍속, 지면 온도에 따른 기상 데이터를 활용하여
안개의 시정 구간을 정확히 예측할 수 있는 머신 러닝 모델을 개발

데이터 분석 프로세스

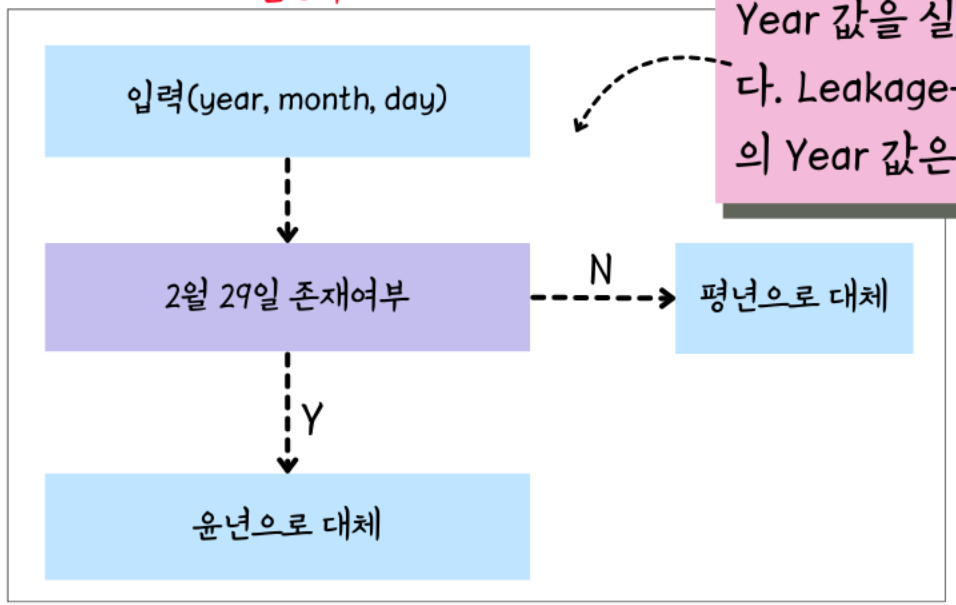

전처리
Raw Data
시간 데이터 처리
결측치 처리 및 보간
Feature Engineering 및 변수 처리
Processed Data
Year 변수 치환
datetime 열 추가
결측치 대체
결측치 보간
Feature Engineering
불필요 변수 제거
모델 학습
Bad
데이터 파이프라인 튜닝
Train Data
Validation Data
모델링 및 평가
데이터 파이프라인 확정
Good
모델 학습
최종 모델
Prediction
Train Data
Test Data

시간데이터 처리 & 결측치 처리

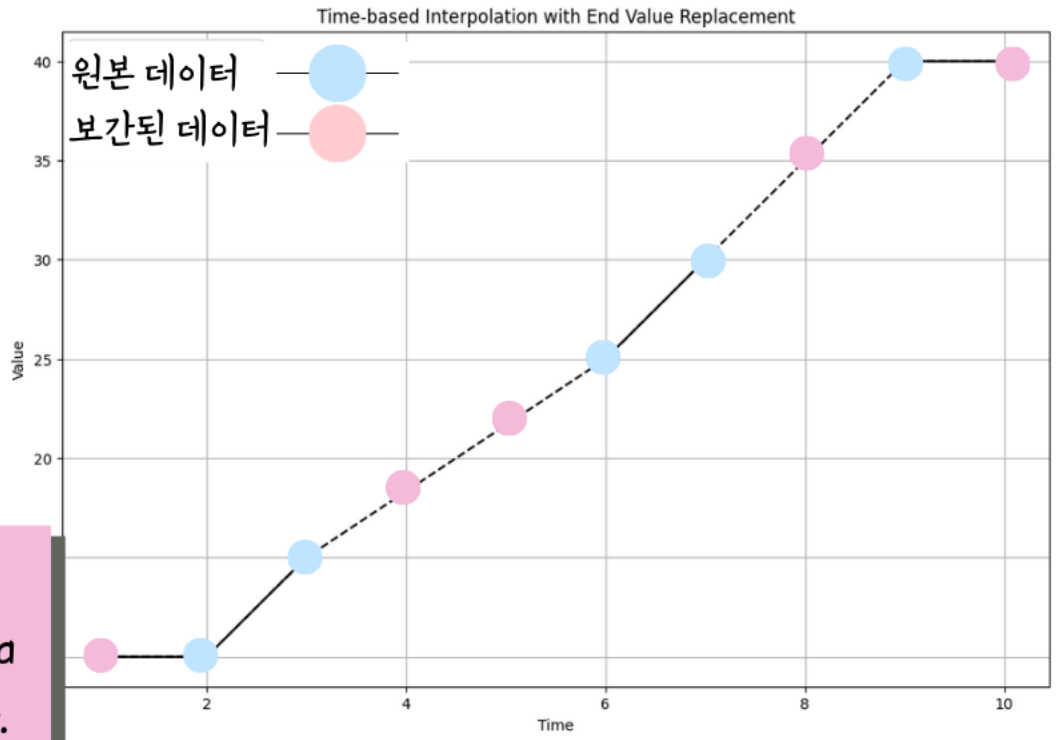


오전 6시

주기적인 패턴을 반영하기 위해 datetime 열을 추가했다.



Year 값을 실제 연도로 매핑해서 사용했다. Leakage를 방지하기 위해, test data의 Year 값은 함수를 사용하여 매핑했다.



데이터의 결측치(약 5.1%)를 그룹별로 보간하였다. 시간의 연속성과 패턴을 반영하기 위해 시간 선형보간법을 수행하고, 양 끝 결측치에 대해서는 forward fill, backward fill을 사용하여 결측치를 처리하였다.

Feature Engineering

'Temp_diff'
지표와 공기 간 온도차

복사안개는 지표면의 복사냉각으로 인해 발생하는 안개의 종류 중 하나로, 바람이 거의 없고 상대습도가 90% 이상으로 높을 때 지표의 온도가 공기의 온도보다 낮아지면 발생한다. 이를 반영하기 위해 Temp_diff 변수를 생성하였다.

$$Temp\_Diff = 기온(ta) - 지면온도(ts)$$

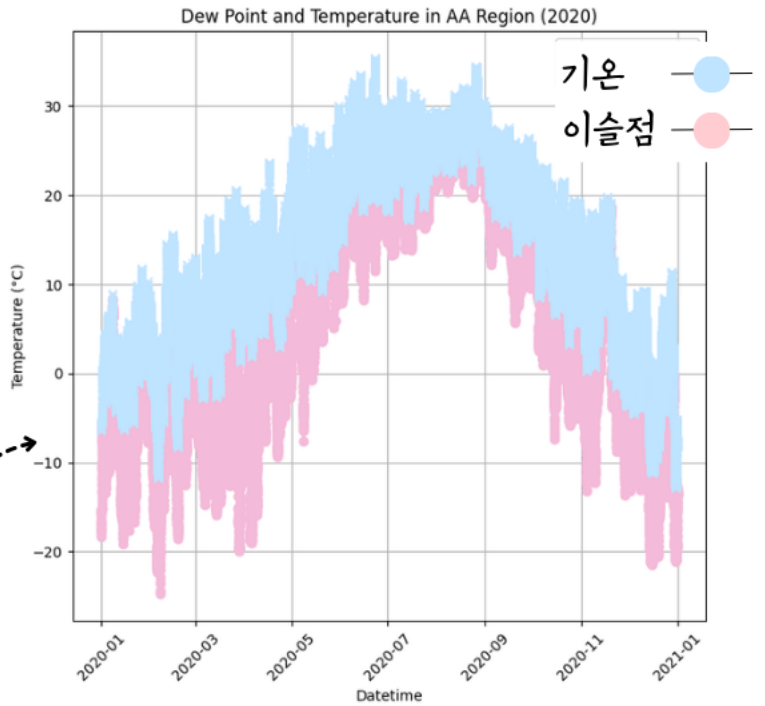
'dew_point' 'dew_point_minus_ta'
이슬점과 기온의 차이

이슬점은 공기가 포화 상태에 도달하여 응결이 시작되는 온도로, 기온이 이슬점 이하로 떨어지면 응결이 발생한다. 이러한 관계를 반영하여, 이슬점과 기온의 차이를 나타내는dew_point_minus_ta 변수를 추가하였다.

기온이 이슬점 이하로 떨어지면 응결이 발생하고 이 과정에서 기화열이 방출되어 기온을 높이므로 기온은 항상 이슬점보다 높을 수밖에 없다. 데이터에서도 이러한 관계를 확인할 수 있었다.

$$dew\_point\_minus\_ta = -\left(이슬점(dew\_point) - 기온(ta)\right)$$

이슬점과 기온



마그누스 공식

$$a = 17.62$$

$$b = 243.12$$

$$\gamma = \frac{a \cdot temperature}{b + temperature} + ln\ ln\left(\frac{humidity}{100.0}\right)$$

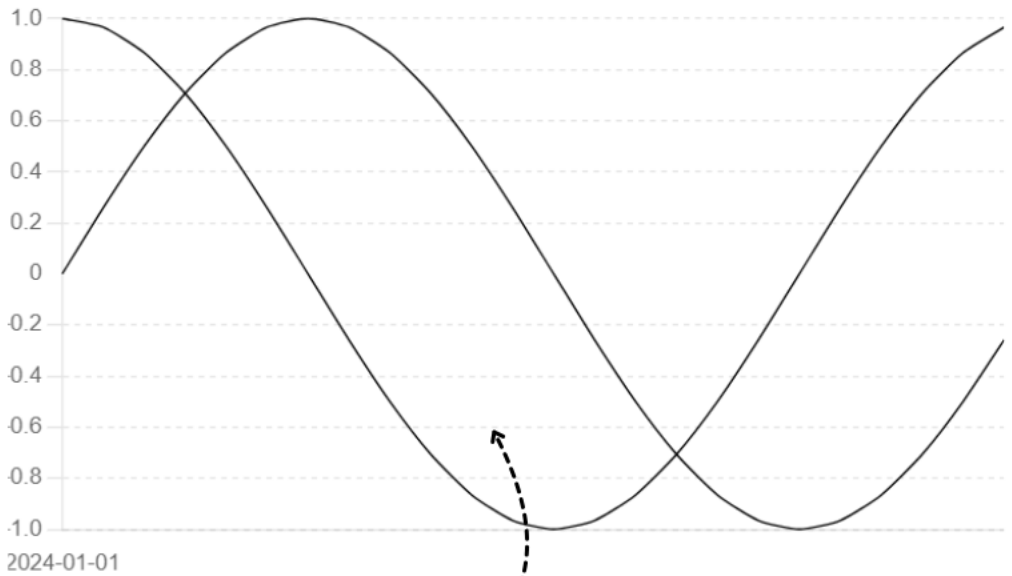
$$dew\_point = \frac{b \cdot \gamma}{a - \gamma}$$

Feature Engineering

'seconds_in_day', 'seconds_in_year'
연초부터 경과 시간/하루 내의 시간 초 단위 변환

'time_sin', 'time_cos'
하루 내의 시간 삼각함수 변환 값

'yearly_time_sin', 'yearly_time_cos'
연초부터 경과 시간 삼각함수 변환 값



모델이 데이터의 주기적 특성을 더 잘 학습할 수 있도록, 연초부터 경과 시간과 하루 내의 시간을 초 단위로 변환한 후, 이를 sine, cosine인 값으로 변환하였다.

'ws10_deg_sin', 'ws10_deg_cos'
풍향의 삼각함수 변환 값

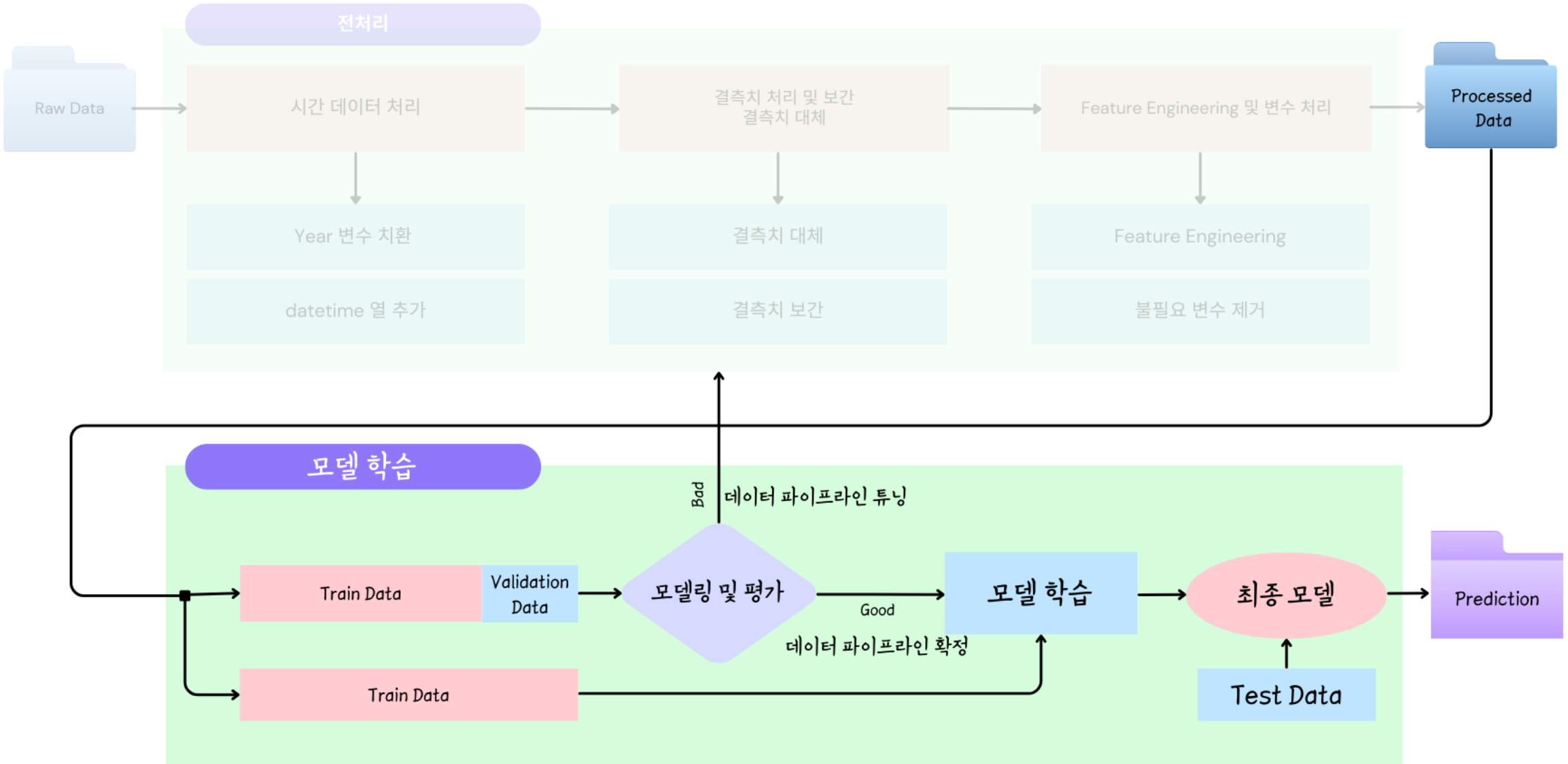
원본 데이터에서 359도와 1도는 매우 다른 값으로 나타나지만, 실제로는방향 차이가 거의 없다는 문제를 삼각함수 변환을 통해 해결할 수 있다. $\sin(359.9°) \fallingdotseq \sin(0.1°) \approx 0$

'rolling_std', 'mean'
특정 기간의 이동 평균/이동 표준편차

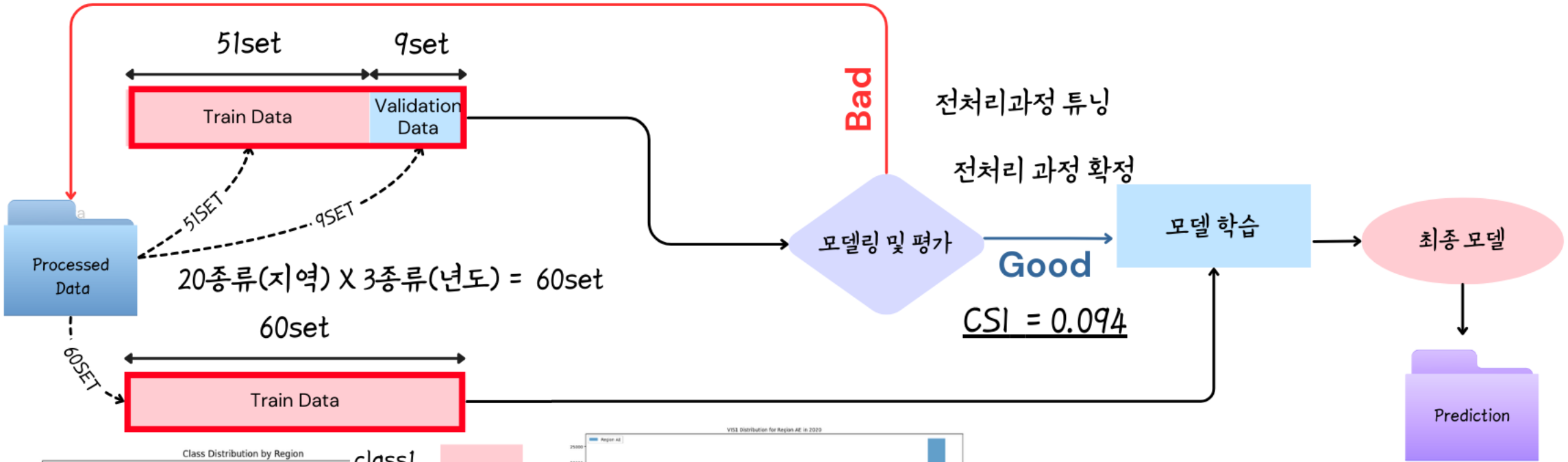
기상 관측값의 높고 낮음을 나타내는 기상 관측값의 이동평균을 시계열 특성으로 추가하는 대신 이동표준편차나 기상 관측값의 상대적인 차이값의 이동평균을 추가하였다.

안개 발생은 응결핵, 풍향, 풍속, 기온의 역전 등 다양한 요인들이 영향을 미치기 때문에, 단순히 기상 관측값이 높거나 낮다고 해서 발생하지 않는다. 지역별로 안개가 발생하기 위한 조건이 다르다. 또한 지역별로 안개가 많이 발생하는 계절이 다르다.

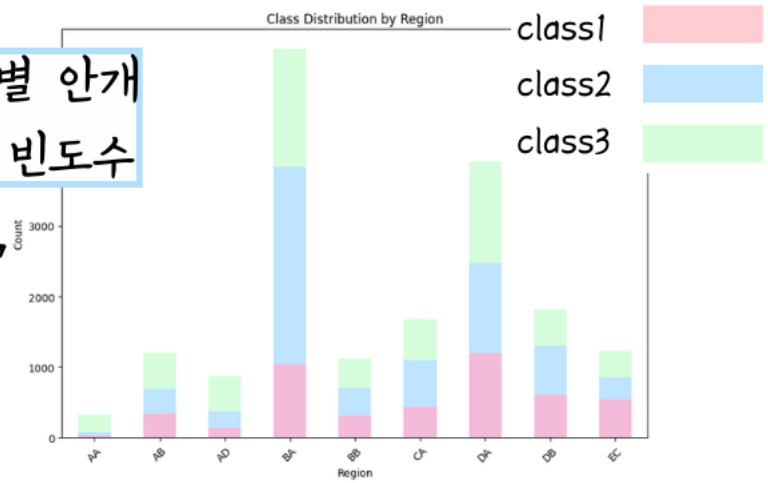
데이터 분석 프로세스


전처리
Raw Data
시간 데이터 처리
결측치 처리 및 보간
결측치 대체
Feature Engineering 및 변수 처리
Processed Data
Year 변수 치환
결측치 대체
Feature Engineering
datetime 열 추가
결측치 보간
불필요 변수 제거
모델 학습
Bad
데이터 파이프라인 튜닝
Train Data
Validation Data
모델링 및 평가
Good
모델 학습
최종 모델
Prediction
데이터 파이프라인 확정
Train Data
Test Data

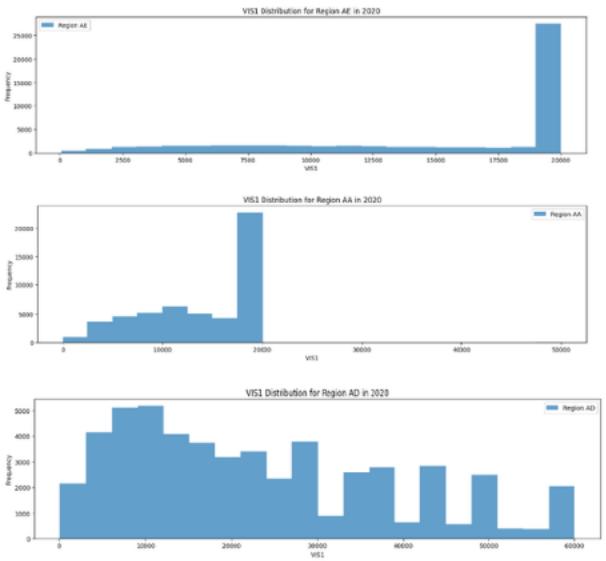
모델 학습및 예측



지역별 안개 발생 빈도수



다양한 분포의 데이터를 validation set으로 사용



각 지역마다 시정(Vis1)의 분포와 스케일 차이로 인해 Regression 이 아닌 Classfication을 사용하기로 하였음.

2020년 AE, AA, AD 지역의 시정 분포(x축: 시정, y축 빈도)

모델상세

모델 파이프라인

데이터 준비

특징과 라벨을 분리한다.
라벨 값을 0부터 시작하도록 조정한다
불균형 데이터를 처리하기 위해 클래스 가중치.

DMatrix 생성

훈련 데이터와 테스트 데이터를 위해
DMatrix 객체를 생성한다

모델학습

XGBoost 파라미터 정의
num round =1100

예측

테스트 데이터에 대한 예측을 수행한다.
원래 라벨 형식과 일치하도록 예측 값을 조정한다

모델 평가

특징 중요도를 계산하고 출력한다.
예측 값의 비율을 계산하고 출력한다

최종 예측 결과

CSI 0.131

*다중 CSI 지표

$$CSI = \frac{H}{H+F+M}$$

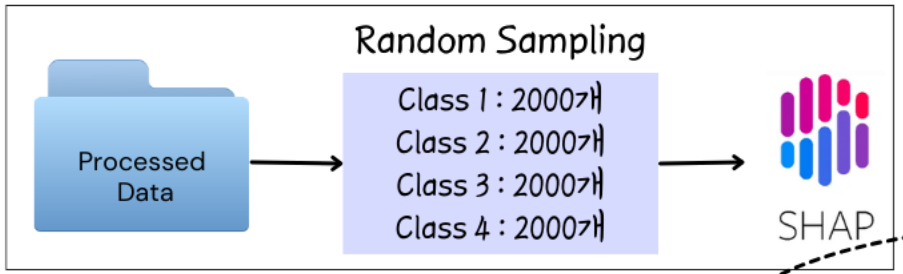
CSI= H / (H+M+F)
다수 값인 class 4의 TP값은 제외 되어 상대적으로 CSI값이 낮게 나온다

XGBOOST의 장점

XGBoost

처리 속도가 빠르다.
가용 메모리가 제한적인 상황에서도 효율적인 메모리 사용이 가능하다.
시계열 데이터 특성상 오버샘플링이 불가능한 상황에서, class별 가중치를 적용하기에 적합하다

분석1. SHAP & Feature Imporatance

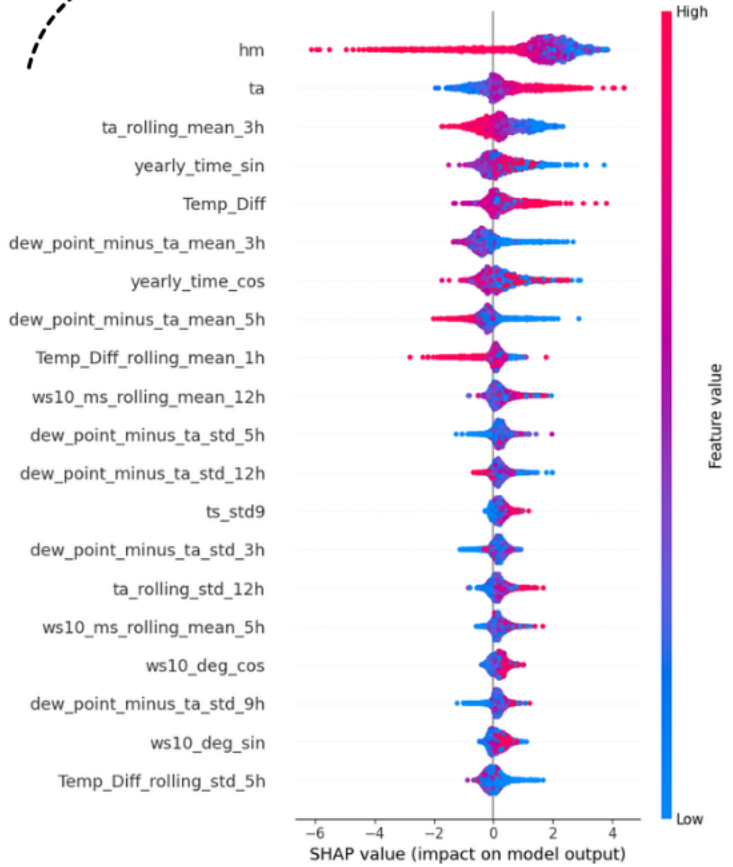


주요 feature

상대습도(hm)

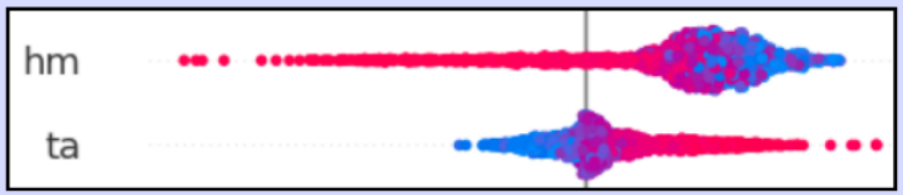
기온(ta)

이슬점과 기온의 차이 이동평균 (dew_point_minus_ta_rolling_mean)



Class1(좌) Class4(우) 의 SHAP Plot(Class2와 Class3은 생략)

Class4의 일부분



높은 기온과 낮은 습도는 시정을 개선하고, 낮은 기온과 높은 습도는 시정을 악화시키는 것으로 나타났다. 안개가 주로 새벽에 발생한다는 것과 주로 새벽의 온도가 낮의 온도보다 낮다는 것으로 설명될 수 있다.

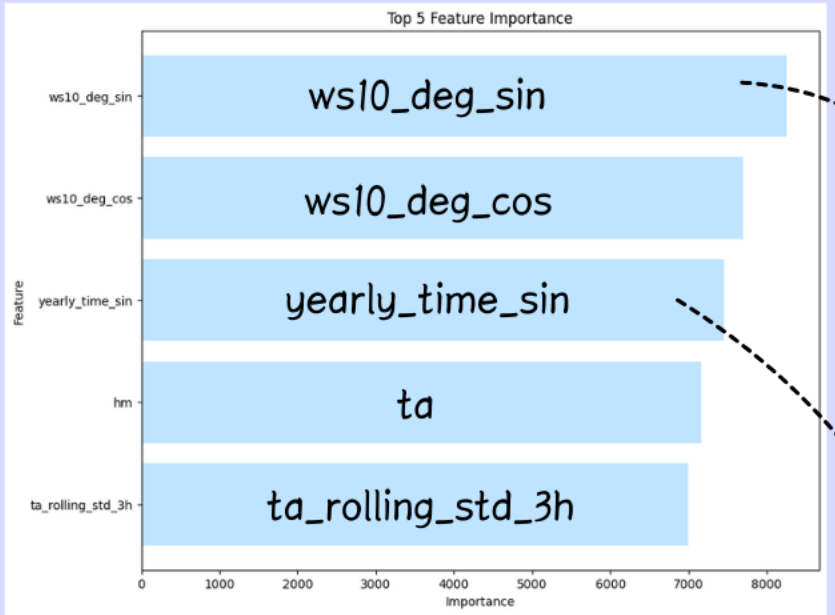
Class별 'hm', 'ta' Shapely value 순위표

	시정구간 1	시정구간 2	시정구간 3	시정구간 4
hm (상대습도)	5위	1위	1위	1위
ta (기온)	1위	3위	7위	2위

안개 발생 시 시정 구간1 인 경우에는 ta(기온)의 영향력이 가장 큰 반면 다른 시정구간에서는 hm(습도)의 영향이 가장 크게 나타났다. 시정 구간에 따라 영향력이 큰 기상관측 변수가 차이가 있음을 확인하였다.

분석1. SHAP & Feature Imporatance

Feature Importance TOP5



SHAP 분석과 달리 ws10_deg_sin, ws10_deg_cos 등 값의 크기가 안개 발생과 관련성이 상대적으로 적었던 특성들이 높은 중요도도를 나타냈다.

ws10_deg_sin, ws10_deg_cos

바람 방향이 안개 형성과 소멸에 영향을 미칠 수 있기 때문에 중요하게 나타난 것으로 해석할 수 있다.

지역(20개)마다 안개 발생 빈도가 다르며, 특정 풍향이 우세하여 풍향의 분포에 차이가 있다 . 이러한 이유로 모델 학습 시 풍향이 안개 발생 빈도를 예측하는 중요한 변수로 작용하여 feature importance가 높게 나타난다.

yearly_time_sin

기장관측값(hm, ta)의 크기와 안개 발생 정도가 선형적으로 관계있는 것이 아니라, 계절마다 안개 발생 정도가 다르고, 특히 특정 월에 안개 발생이 많기 때문에 중요한 피처로 나타난 것이다.

분석2. 지역별 기상 관측 변수의 통계적 차이 분석

분석 가설

지역 특성별로 발생하는 안개가 다르므로, 기존 모델에는 A, B,C, D, E (stn_id의 앞글자) 지역 특성 변수가 예측에 있어 중요할것이다.

feature importance 분석 결과

지역 특성 변수 A, B,C, D, E가 항상 feature importance에서 최하위에 위치했다. 지역특성 자체보다는 다른 기상 변수들이 안개 발생 예측에 더 중요한 영향을 미친다.

클러스터링 분석

안개 발생(class 1, 2, 3) 데이터 중 'ws10_ms', 'ta', 'ts', 'sun10', 'hm' 에서 가능한 모든 조합(31개)에 대해 K-means 클러스터링을 수행하고, 클러스터링 결과를 바탕으로 지역 특성 변수 A, B,C, D, E로 잘 그룹화 되는지 Adjusted Rand Index (ARI)를 통해 평가했다.

클러스터링 분석 결과

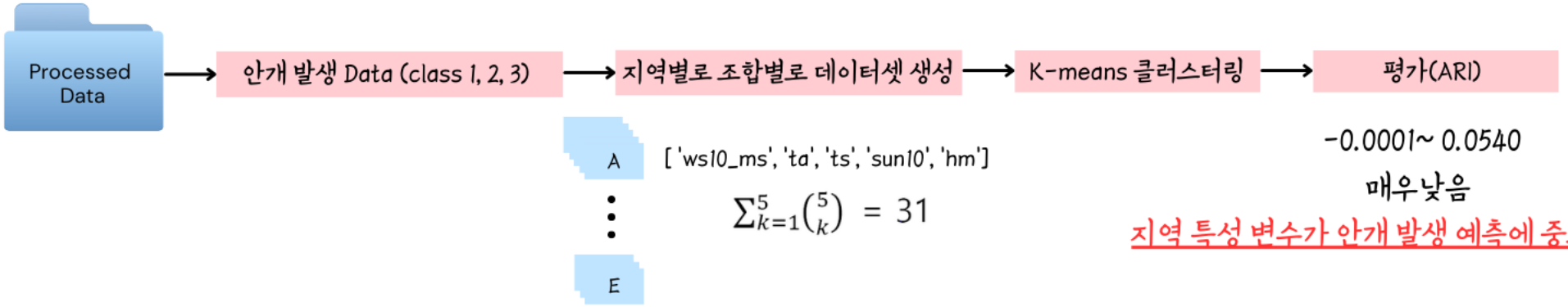
조합 서른 한 개에 대해 ARI값은-0.0001~ 0.0540으로 낮게 나타났다. 지역 특성 변수 A, B,C, D, E가 각 지역의 특성을 잘 반영하지 못함을 시사한다. 즉, 지역 특성 변수가 실제로 안개 발생 예측에 중요하지 않음을 알 수 있다.

Adjusted Rand Index (ARI)

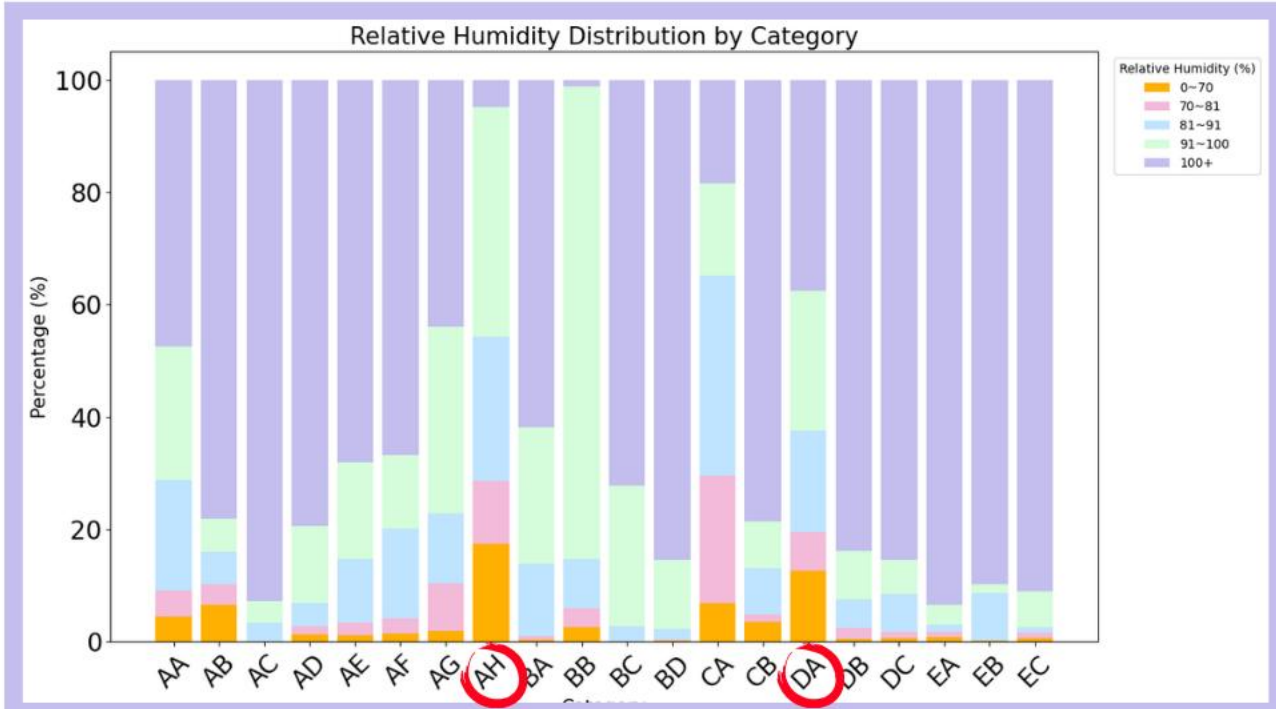
$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]}$$

(RI : Rand Index, $E[\text{RI}]$: Expected Rand Index)

클러스터링 결과와 실제 지역 특성 간의 일치도를 평가하는 지표로, 1에 가까울수록 클러스터링이 잘 된 것을 의미한다.



분석3. 상대습도 구간별 안개발생 비율



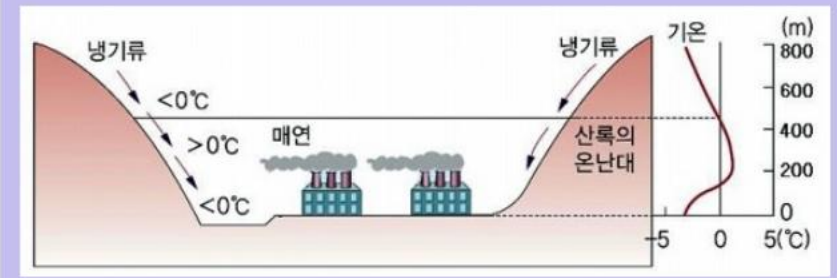
상대습도가 91% 이상일 때 대부분의 안개가 발생했다.
그러나 AH, DA 지역에서는 상대습도70% 이하 구간에서도
다른 지역에 비해 안개 발생 비율이 현저히 높게 나타났다.

→원인 분석

공장지대

상대습도가 낮더라도 해면에서 대기로 날아온 염분을 띤 작은 결정, 굴뚝 연기의 입자 등 흡습성 미립자의 포함량이 많은 경우 안개 발생이 증가할 수 있다.

고도에 따른 기온 역전 현상



고도에 따른 기온 역전 현상이 심한 경우 안개 발생이 증가할 수 있다.

활용 방안및 기대효과

Key Findings

Pg.13 AH와 DA에서는 예외적으로 70%이하의 습도에서 안개 발생 비율이 높았다. 주어진 데이터 외 다른 요인들도 고려해야 한다.

Pg.12 지역 특성 변수 A, B,C, D, E가 각 지역의 특성을 잘 반영하지 못한다. 즉, 지역 특성 변수가 실제로 안개 발생 예측에 중요하지 않다.

Pg.7 단순히 기상 관측값이 높거나 낮다고 해서 발생하지 않는다. 지역별로 안개가 발생하기 위한 조건이 다 다르다. 또한 지역별로 안개가 많이 발생하는 계절이 다르다.

Pg.10 안개 발생 시 시정 구간1 인 경우에는 ta(기온)의 영향력이 가장 큰 반면 다른 시정구간에서는 hm(습도)의 영향이 가장 크게 나타났다. 시정 구간에 따라 영향력이 큰 기상 관측 변수가 차이가 있음을 확인하였다.

시사점

고도에 따른 기온 정보와 주변 지역 정보(예: 공장 지대)를 포함한 지역 관련 데이터를 수집하여 데이터셋을 구축하는 등 지역 특성을 잘 반영하는 특징을 모델에 추가한다면,더 정확한 예측을 수행할 수 있을 것이다.

기존의 지역 분류(A, B, C, D, E)이 아닌 지역 특성을 더 잘 반영할 수 있도록 20개의 지역을 클러스터링하여 지역특성을 다시 정의한다면 모델을 성능을 개선할 수 있을 것이다.(혹은 지역별로 예측모델 제작).

안개 발생 시, 시정 구간 1에서는 기온 변화를 중심으로 안개 해소 가능성을 평가하고, 다른 시정 구간에서는 습도를 집중 모니터링하여 안개 발생(안개 해소 시점, 지속시간)을 예측해야한다.

다양한 기상정보

주변 지역 정보

새롭게 정의된 지역특성별 모델

or

지역별 예측모델

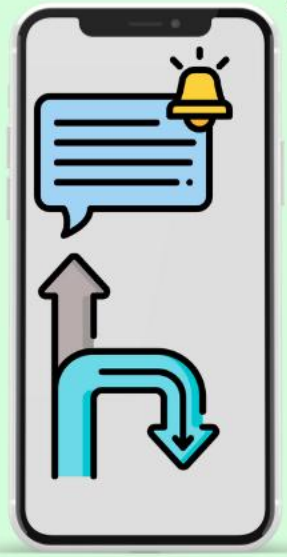
모델 성능 향상

시정 구간에 따른 모니터링 방법 변화

활용 방안및 기대효과

모델 성능 향상

시정 구간에 따른
모니터링 방법 변화



교통 관리 시스템

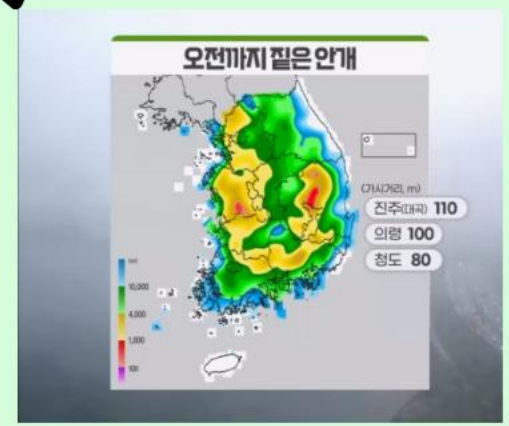
안개발생시 시정구간 별로 모니터링 방법을 달리하여 안개 해소 시점을 운전자에게 알리고, 이를통해 사고발생 감소시킬 수 있다.

운전자 도로 진입 전, 안개 발생이 예상되는 도로를 우회할 수 있도록 운전자에게 정보를 제공하여 교통체증및 사고발생 방지.

항공 교통 관제 시스템

항공기 이착륙 지연 시간을 정확하게 예측하여, 승객들에게 신속하고 정확한 정보를 제공함으로써 불편을 최소화할 수 있다.

안개 발생 시점을 정확하게 예측하여 항공 사고를 예방할 수 있다.



기상 예보 서비스 품질 향상

안개 발생 예측 AI 모델의 성능을 향상시킴으로써 예보 시스템을 강화할 수 있다.

안개발생시 시정구간 별로 모니터링 방법을 달리하여 안개해소 시점을 정확하게 예측하고 예보 서비스의 신뢰성을 향상시킬 수 있다.