시계열 특성과 balanced weight을 사용한 기상 특성에 따른 안개 발생진단 최종 CSI:0.131

참가번호	240184	팀명	근두운

과제2- 기상특성에 따른 안개 발생 진단

1. 배경 및 목표

안개는 항공기 사고, 교통 지체, 호흡기 문제와 같이 우리 일상과 밀접하게 연결되어 있다. 내륙, 내륙산악, 동해안, 서해안, 남해안의 5개 지역에서 각각 다른 특성을 보이며, 활승안개, 복사 안개, 이류 안개 등 다양한 종류가 존재한다. 이러한 복잡성을 고려하여, 본 공모전의 목표는 시간에 따른 습도, 풍속, 지면 온도에 따른 기상 데이터를 분석하여 발생한 안개의 시정 구간을 정확히 예측할 수 있는 머신 러닝 모델을 개발하는 것이다. 이를 통해 다양한 지역적 특성을 반영한 예측 모델을 구축하고, 기상 예측의 정밀도를 높이고자 한다.

2. 분석 데이터 정의

항목	내용
데이터 제공	기상청 플랫폼 날씨마루
데이터 수집 방법	dbGetQuery사용하여 다운로드
데이터 형식	CSV
데이터셋의 크기	(3156459, 16)
데이터 범위와 기간	Train data = 3년, Test data = 1년(1월1일~12월 31일 10분단위)

	Value	Count	Frequency (%)
0	4.0	3101809.0	98.3
1	-99.0	22516.0	0.7
2	3.0	12180.0	0.4
3	2.0	12088.0	0.4
4	1.0	7866.0	0.2

시정 구간 4 에 해당하는 데이터가 전체 데이터의 98.3%를 차지하고 있어 타겟 레이블이 불균형함을 확인하였다. 또한, -99.0 으로 표시된 값은 결측치로 가정하였다.

3. 데이터 분석, 전처리 및 파생 변수 생성

3.1. year변수 치환

시계열 특성을 이용하기위해 알파벳으로 되어 있는 year값을 실제 년도로 치환해서 사용하였다. 이때 leakage 방지를 위해, test데이터에서는 함수를 사용하여 대체한다. 시계열 모델 구축 과정에서 제공된 데이터를 가공하여 파생 변수를 생성했다. 가공하여 추가한 변수는 3 절에서 확인할 수 있다.

3.2. 결측치 처리(interpolate: time)

먼저 데이터의 존재하는 결측값 -99, -99.9를 NaN으로 처리하였다. 주어진 데이터는 시계열 데이터 이기에, 시간적 연속성을 유지하는 것이 중요하다. 따라서 시간에 따른 연속적인 변화를 나타내기 위해 결측값에 대하여 시간 선형 보간**(interpolate: time)**을 수행하였다. 먼저, interpolate_and_fill 함수로 vis1 값을 채운 뒤, 양 끝의 결측치는 앞뒤 값으로 채우고, fill_class 함수를 사용해 class 열의 결측치를 vis1 값을 기준으로 채운다. 이후, vis1과 class를 제외한 나머지 수치형 변수들의 결측치를 시계열 보간 방법으로 채우고, 양 끝의 결측치는 앞뒤 값으로 채운다. 이 때 'stn_id' 기준으로 그룹화한 후, 각 그룹별로 결측치를 보간한다.

3.3. feature_engineering

'Temp_diff' : 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('ta' – 'ts')를 파생변수로 추가하였다. Temp_diff는 기온(ta)에서 지면온도(ts)를 뺀 값으로, 지표와 공기 간의 온도 차이를 나타낸다.

$$Temp\_Diff = 기온(ta) - 지면온도(ts)$$

'dew_point', 'dew_point_minus_ta': 이슬점은 공기가 포화 상태에 도달하여 응결이 시작되는 온도로, 기온이 이슬점 이하로 떨어지면 응결이 발생하고 이 과정에서 기화열이 방출되어 기온을 높인

다. 이슬점 계산을 위해 마그누스 공식(Magnus formula)을 사용하였다. 공식에 따르면, 기온은 항상 이슬점보다 높을 수밖에 없고 데이터에서도 이러한 관계를 확인할 수 있었다. 이를 반영하여, 이슬점과 기온의 차이를 나타내는 dew_point_minus_ta 변수를 추가하였다. 각 변수에대한 계산 식은 다음과 같다.

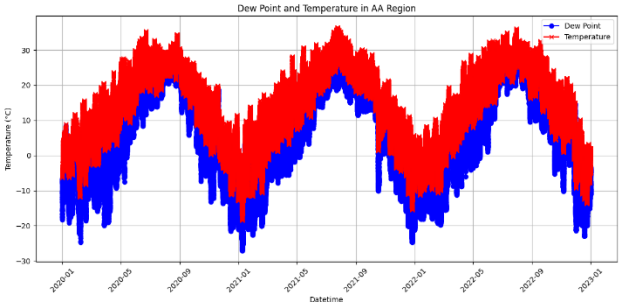
$$a = 17.62$$

$$b = 243.12$$

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

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

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



이슬점과 기온 예시(AA지역)

3.4. 시계열 특성 설명(rolling_std, mean)

안개 발생은 단순히 기상 관측값이 높거나 낮다고 해서 발생하지 않는다. 이는 응결핵, 풍향, 풍속, 기온의 역전 등 다양한 요인들이 안개 발생에 영향을 미치기 때문이다. 예를 들어, 흡습성 미립자와 바람 조건, 기온 역전 등이 복합적으로 작용하여 안개가 형성된다. 따라서 안개 발생은 단순히 기상 관측값이 높거나 낮다고 해서 발생하지 않는다. 지역별로 안개가 발생하기 위한 조건이 다 다르기 때문이다. 여러 기상 관측값 및 시간값들의 조합이 지역별 안개 발생 조건과 일치할 경우 안개가 발생한다.

이러한 이유로 기온(ta)과 풍속(ws10_ms)을 제외하고, 기상 관측값의 높고 낮음을 나타내는 기상 관측값의 이동평균을 시계열 특성으로 추가하는 대신 이동표준편차나 기상 관측값의 상대적인 차이값의 이동평균을 추가하였다. 기온(ta)과 풍속(ws10_ms)의 경우 이동평균을 추가한 이유는 기온(ta)의 경우 모델 제작 시 feature importance가 항상 높게 나왔기 때문이며, 풍속(ws10_ms)은 안개 발생과 인과관계가 분명하며 이동표준편차의 의미가 상대적으로 적을 것으로 예상했기 때문이다

3.5. 삼각함수 변환

사용된 기상 데이터의 거의 모든 feautre들이 주기성을 갖고있음을 시각적으로, 자기상관을 측정하여 확인하였다. 또한 주기적인 패턴을 효과적으로 처리하기 위해 삼각함수 변환을 사용하였다. 풍향 변환에서는 각도를 라디안 단위로 변환한 후 사인과 코사인 값을 계산하여 방향성을 주기적으로 표현하였고, 시계열 데이터 변환에서는 하루 및 연간 시간 정보를 사인과 코사인 값으로 변환하여 시간의 주기성을 반영하였다. 이를 통해 모델이 시간과 방향의 주기적 특성을 더 잘 학습할 수 있게 만들었다.

'ws10_deg_sin', 'ws10_deg_cos' : 풍향(ws10_deg)을 라디안 단위로 변환한 후 각각 사인(sin)과 코사인(cos) 값을 계산한 변수이다. 풍향의 경우, 359도와 1도는 방향 차이가 얼마 나지 않지만, 원본 데이터에서는 이를 반영하기 힘들어 완전히 다른 값이 되어버린다. 따라서 풍향의 주기적인 특성을 반영하기 위해 삼각함수 변수인 (ws10_deg_sin, cos)을 추가하였다.

'seconds_in_day': 하루 내의 시간을 초 단위로 변환한 변수이다. 시간 정보(hour, minute)를 초 단위로 변환하여 하루 주기를 나타낸다.

'***seconds_in_year':*** datetime을 기준으로 연도 시작부터 현재까지의 경과 시간을 나타낸다.

'time_sin', 'time_cos' : 하루 주기의 사인 값과 코사인 값을 계산한 변수이다.

'yearly_time_sin', 'yearly_time_cos' : 1년 주기의 사인 값과 코사인 값을 계산한 변수이다.

3.6. drop한 feature

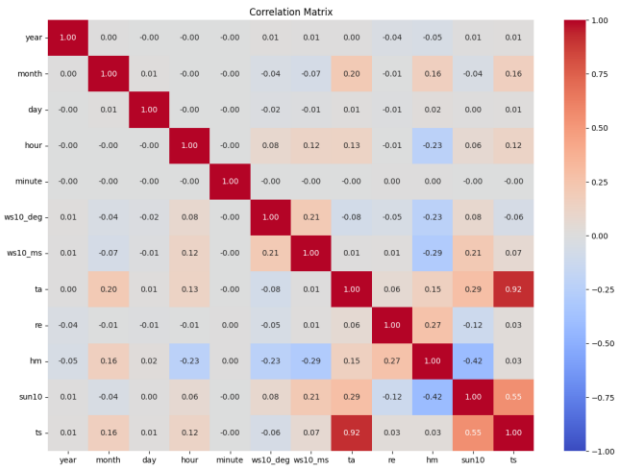
'datetime': 시계열 특성을 추가하기위해 사용된변수로, 보통 학습에 직접적으로 사용되지 않는다.

'month', 'day', 'hour', 'minute', 'seconds_in_day' 'seconds_in_year': 시계열 특성을 추가하기 위해 사용되었고, 삼각함수를 사용하여 시계열 특성으로 변환된 변수가 있어 불필요하다.

'dew_point': 'dew_point_minus_ta' 추가에 사용되었다.

'dew_point_minus_ta': 모델 제작 시 feature importance가 항상 하위권에 존재하여 사용하지 않았다. 그러나 변수의미를 반영하기위해 이동평균, 이동표준편차 특성을 추가하기위해 사용되었다.

'ts': 결측치 보간 후 변수간 상관계수를 파악하기 위해 heatmap으로 상관계수를 확인해본 결과, ta와 ts 간의 피어슨 상관계수가 0.92로, 다중공선성이 의심되었다.



ta와 ts 변수간의 VIF값이 약 16.445로 10을 초과하여 ta와 ts가 다중공선성이 높음을 확인하였다. 그러나 지표면온도(ts)는 안개 발생 예측의 중요한 변수라고 생각되어 파생변수 Temp_Diff 제작에 사용한 후, drop하였다.

're', 'sun10': 모델 제작 시 feature importance가 현저히 낮아, 사용하지 않았다.

'ws10_deg': 각도 데이터는 0도와 360도가 같은 위치를 가리키는 주기적 특성을 가지고 있는데, 이를 고려하지 않고 단순히 각도 값을 사용하면 0도와 360도 사이에 큰 차이가 있는 것처럼 보이기 때문에, 이러한 주기성과 연속성을 표현하기 위해 sine 과 cosine으로 변환하고 drop하였다.

stdn_id: 지역 특성별로 발생하는 안개가 다르므로, 기존 모델에는 A, B, C, D, E(stdn_id의 앞글자) 지역 특성 변수가 예측에 있어 중요할 것으로 예상하였고 이를 모델 학습에 포함시켰지만, 항상 feature importance에서 최하위에 위치했다. 이에 대한 이유를 분석하기 위해, 다양한 기상 관측 변수 조합을 사용하여 클러스터링을 수행하고 Adjusted Rand Index (ARI)를 계산하였다.

3.7. 지역 특성별 기상관측 변수의 통계적 차이 분석

지역 특성별로 발생하는 안개가 다르므로, 기존 모델에는 A, B, C, D, E (stn_id의 앞글자) 지역 특성 변수가 예측에 있어 중요할 것으로 예상했다. 이를 모델 학습에 포함시켰지만, 항상 feature importance에서 최하위에 위치했다. 이에 대한 이유를 분석하기 위해, 다양한 기상 관측 변수 조합을 사용하여 클러스터링을 수행하고 Adjusted Rand Index (ARI)를 계산했다. ARI는 클러스터링 결과와 실제 지역 특성 간의 일치도를 평가하는 지표로, 1에 가까울수록 클러스터링이 잘 된 것을 의미한다. 다음은 ARI의 식이다.

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

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

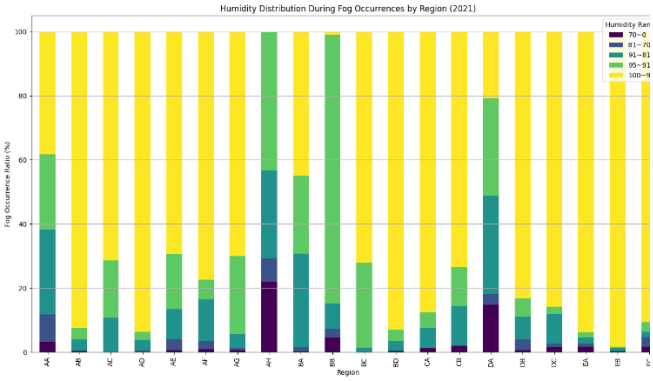
데이터 준비 단계에서는 안개 발생(class 1, 2, 3)을 필터링하여 기상 관측 데이터를 준비했다. 변수 조합 생성 단계에서는 'ws10_ms', 'ta', 'ts', 'sun10', 'hm' 다섯 개의 기상 관측 변수의 가능한 모든 조합을 생성했다. 클러스터링 수행 단계에서는 각 변수 조합에 대해 K-means 클러스터링을 수행하고, 클러스터링 결과를 바탕으로 Adjusted Rand Index (ARI)를 계산했다. 결과 평가 단계에서는 각 조합의 ARI를 평가하여 지역 특성 변수의 중요성을 분석했다.

각 변수 조합에 대해 ARI값의 범위는 -0.0001~ 0.0540으로 낮게 나타났다. 이는 지역 특성(즉, stn_id의 앞글자)에 따른 기상 관측 변수들이 클러스터링에 있어서 큰 영향을 미치지 않음을 시사한다. 즉, 지역 특성 변수가 예측 모델에서 feature importance가 낮은 이유는 이 변수가 실제로 안개 발생 예측에 중요하지 않기 때문일 가능성이 높다. 따라서 stdn_id를 drop하였다.

지역 특성을 클러스터링하여 지역 특성을 새로 정의하려는 시도를 하였으나, 한정된 시간으로 진행하지 못했다. 향후 이러한 분석을 완료하면 모델을 개선할 수 있을 것으로 예상한다.

3.8. 상대습도 구간별 안개발생 비율

10분 평균 상대습도(%)	70~0	81~70	91~81	95~91	100~95
AA	4.41	4.75	19.66	23.73	47.46
AB	6.53	3.64	5.87	5.87	78.1
AC	0	0	3.35	3.91	92.74
AD	1.25	1.48	4.22	13.68	79.36
AE	1.16	2.24	11.3	17.18	68.11
AF	1.49	2.7	15.95	13.11	66.76
AG	1.99	8.47	12.35	33.17	44.02
AH	17.37	11.28	25.66	40.82	4.87
BA	0.38	0.65	12.91	24.31	61.74
BB	2.65	3.2	8.87	84.1	1.19
BC	0.04	0.04	2.63	25.1	72.18
BD	0.13	0.23	1.96	12.19	85.5
CA	6.88	22.74	35.52	16.41	18.46
CB	3.57	1.19	8.33	8.33	78.57
DA	12.63	6.79	18.04	24.98	37.56
DB	0.49	1.97	4.99	8.72	83.82
DC	0.58	1.15	6.67	6.1	85.5
EA	0.8	1	1.2	3.59	93.41
EB	0.1	0.1	8.37	1.69	89.73
EC	0.65	0.98	0.9	6.37	91.09



상대습도 구간별 안개발생 비율 표와 그림(%)

높은 상대습도는 이슬점을 높이고 안개 발생 확률을 증가시킨다. 위의 표와 그래프는 상대습도 구간별로 안개 발생(class: 1, 2, 3)이 얼마나 자주 발생하는지 백분율로 계산한 결과이다. 분석 결과, 상대습도가 91% 이상일 때 대부분의 안개가 발생했다.

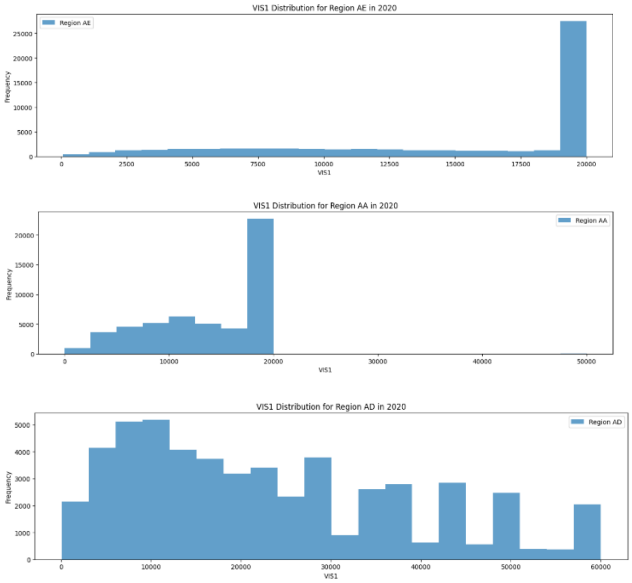
그러나 AH, DA 지역에서는 70% 이하 구간에서도 다른 지역에 비해 안개 발생 비율이 현저히 높게 나타났다. 이는 상대습도 외에도 해면에서 대기로 날아온 염분을 띤 작은 결정, 굴뚝 연기의 입자 등 흡습성 미립자의 포함량이 많거나, 고도에 따른 기온 역전 현상이 심하기 때문일 것으로 추정된다. 풍속은 오히려 해당 지역(AH, DA)에서 낮은 편임을 확인하였으므로, 풍속은 원인이 아닐 것으로 예측할 수 있다. 지역 특성을 잘 반영하는 특징을 모델에 추가하는 것이 정확한 예측을 위해 중요할 것으로 생각하였다.

만약 고도에 따른 기온 정보나, 주변에 공장지대가 있는지 여부가 주어진다면 더욱 정확하고 신뢰할 수 있는 예측이 가능할 것이다.

4. 모델 구축

Classfication이 아닌 Regression은 시정(Vis1)의 값을 예측하고 이를 바탕으로 시정 구간을 도출할 수 있기 때문에 더 세밀한 예측이 가능하다. 예를 들어, Regression 모델은 시정이 199m인지 201m인지 예측할 수 있으며, 이는 분류 모델이 구분할 수 없는 디테일을 제공할 수 있다.그러나 각 지역마다 시정(Vis1)의 분포는 Regression을 사용하기에 부적절하다.

아래의 히스토그램은 각 지역별(AE, AA, AD)로 시정 값의 분포를 나타낸다. 예를 들어, Region AA에서는 시정 값이 50000 이상으로 이상치처럼 나타나는 경우가 있으나, AD 지역에서는 50000 이상의 데이터가 많이 포함되어 있다. 또한, AE 지역은 20000 이하의 데이터만 포함되어 있다. 이러한 스케일 차이로 인해 Regression 학습을 위해 스케일링하는 과정에서 어려움을 겪었다. 결정적으로는 시계열 데이터의 특성상 결측치가 포함된 데이터를 drop할 수 없기 때문에 대체해야 하는데, 이 과정에서 부정확한 데이터가 발생할 수 있다. 이러한 risk를 줄이기 위해 Classification을 사용하기로 결정하였다.



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

4.1. Class Label 처리 및 가중치 적용

사용할 모델인 XGBoost가 다중 클래스 분류에서 라벨을 0부터 시작하는 것을 요구하기 때문에, 원래 1, 2, 3, 4였던 라벨을 0, 1, 2, 3으로 변환한다. 시계열 데이터 특성상 smote와 같은 oversampling을 사용하기 어렵다. 'compute_class_weight' 함수를 사용하여 각 클래스의 빈도에 따라 가중치를 자동으로 계산하고, 이를 dictionary 형태로 저장한 후, y_train의 각 라벨에 적용하였다. 빠른 처리속도와, 최소표본화 된 클래스 문제를 해결하기 위해 가중치가 적용된 d-matrix를 사용하였다.

4.2. XGBOOST, CSI 계산

XGBoost는 그라디언트 부스팅 프레임워크로, 다중 클래스 분류 문제에서 탁월한 성능을 발휘하며, 클래스 불균형 문제를 해결하기 위한 가중치 적용 기능도 제공한다. 모델 학습을 위해 1100회의 부스팅 라운드를 설정하고, 조기 종료를 위해 검증 데이터를 사용하였다.

```
# XGBoost 모델 학습

param = {
    'objective': 'multi:softmax',
    'num_class': 4,
    'eval_metric': 'mlogloss',
    'eta': 0.3,
    'max_depth': 6,
    'seed': 42
}

num_round = 1100
evallist = [(dtrain, 'train')]
bst = xgb.train(param, dtrain, num_round, evallist, early_stopping_rounds=10)
```

일반적인 classification 지표인 f1-score나 accuracy가 아니기

에, 대회의 평가지표인 CSI값을 계산하기위해 class label을 다시 원래대로 하고, 혼동행렬을 기준으로 CSI값을 계산한다.

```
# 다중 CSI 계산
H = cm[0, 0] + cm[1, 1] + cm[2, 2]  # H11, H22, H33
F = (cm[0, 1] + cm[0, 2] + cm[1, 0] +
     cm[1, 2] + cm[2, 0] + cm[2, 1] +
     cm[3, 0] + cm[3, 1] + cm[3, 2]) # F12, F13, F21, F23, F31, F32, F41, F42, F43
M = cm[0, 3] + cm[1, 3] + cm[2, 3]  # M14, M24, M34

CSI = H / (H + F + M)
print(f"CSI: {CSI:.2f}")
```

5. 모델 예측 결과

5.1. 검증 결과

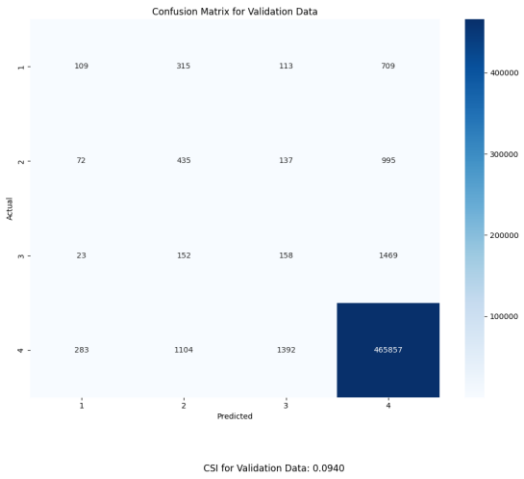
지역별로 Class가 불균형하므로, validation set으로 사용할 지역과 년도를 선정할 때 무작위로 선정할 경우 검증 지표가 모델의 성능을 잘 대변하지 못할 수 있다. 이를 방지하기 위해 각 년도, 지역별로 데이터의 분포를 확인하여, 최대한 다양한 분포의 데이터를 validation set으로 사용하고자 하였다. 최종적으로 다음과 같은 지역과 년도를 선정하였다:

- ***('AA', 2020)***
- ***('EC', 2021)***
- ***('DA', 2022)***
- ***('BB', 2022)***
- ***('CA', 2022)***
- ***('AB', 2021)***
- ***('AD', 2022)***
- ***('BA', 2020)***
- ***('DB', 2021)***

이를 통해, 최대한 다양한 지역의 특성을 반영하여 검증 데이터를 구성함으로써 모델이 다양한 상황에서도 일반화된 성능을 도출할 수 있게끔 하였다.

모델의 예측 성능을 평가한 결과, CSI (Critical Success Index) 값이 약 0.094로 낮게 나타났다. 이는 CSI식과 train_data의 분포와 연관이 있다. 본 대회에서는 실제 안개 발생 예측 성능을 더욱 부각시키기 위해, CSI 계산 시 다수 클래스인 class4의 TP(True positive)값을 계산에서 제외하였다. 다수 클래스가 98.97%을 차지하는 불균형이 심한 분포가 모델이 소수 클래스를 제대로 학습하지 못하게 하여, 예측 성능이 저하되고 CSI 값이 낮게 나오는 결과가 초래된 것이다.

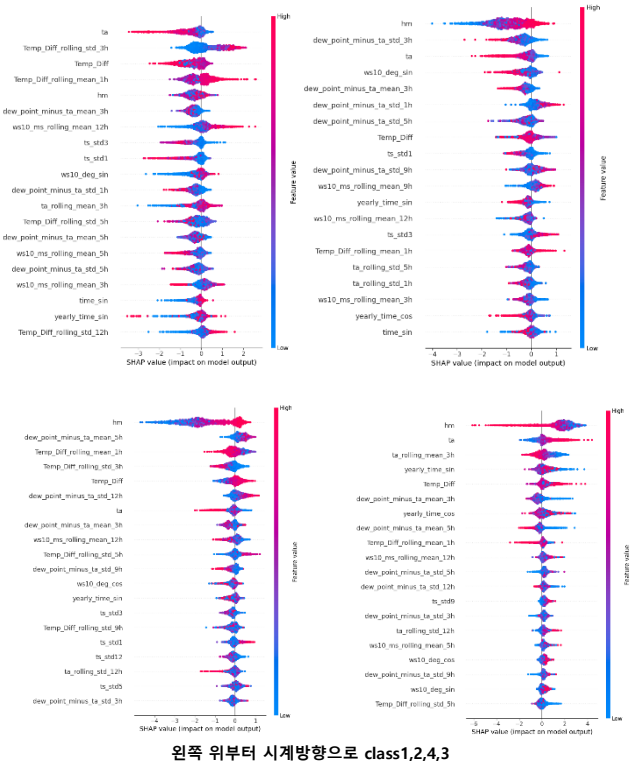
이번 대회에서는 제한된 시간으로 Classification 모델만을 사용하였으나, Class가 1, 2, 3인 데이터만으로 Regression 모델을 제작하여 소수 클래스에 대한 세밀한 예측을 할 수 있을것이다. 먼저 전체 데이터를 Classification 모델을 통해 분류하고, 이후 소수 클래스에 대해 Regression 모델을 적용함으로써, 각 클래스의 특성을 더 세밀하게 예측하는 방식이다. 이를 통해, 소수 클래스에 대한 예측 성능을 향상시키고, 모델의 전반적인 예측 정확도를 높일 수 있을것이다.



5.2. 최종 예측 결과

최종 모델을 valid set을 나누지 않고 전체 train_data에 대해서 학습을 진행하였다. 이는 validation data로 CSI값을 검증하여 파생변수 생성과 모델 fine-tuning을 진행하고, 마지막에 최종 모델을 학습할 때는 전체 train_data를 사용하여 학습에 놓치는 데이터가 없게끔 하였다. 최종 test에 대한 결과로 CSI: 0.131을 도출해냈다.

5.3. 예측 결과 분석



왼쪽 위부터 시계방향으로 class1,2,4,3

SHAP(Shapley Additive exPlanations) 값을 사용하여 기상 데이터의 각 피처가 모델의 예측에 미치는 영향을 평가하였다. 시정(Class)이 낮아질수록 안개가 심해지며 시정이 낮아짐을 의미한다. 또한 각 클래스(1, 2, 3, 4)에 대해 2000개의 샘플을 랜덤으로 추출하여 SHAP 값을 계산하고, 그 결과를 시각화하였다. 다음은 각 클래스에 대한 주요 피처의 해석이다.

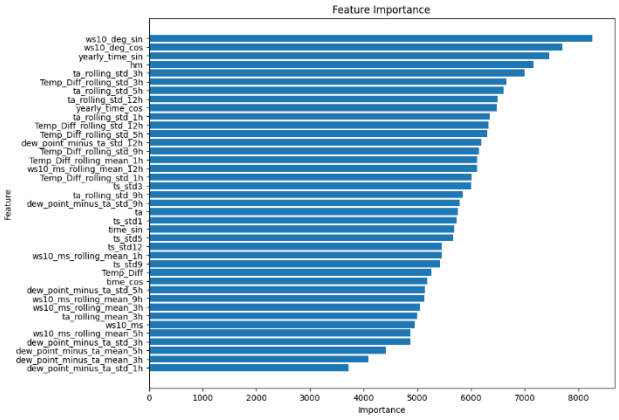
네 개의 클래스 모두에서 상대습도(hm), 기온(ta), 이슬점과 기온의 차이 이동평균(dew_point_minus_ta_rolling_mean)이 중요한 피처임을 확인할 수 있었다. 특히 기온은 모든 클래스에서 시정에 상당히 중요한 영향을 미치며, 높은 기온은 시정을 개선하고, 낮은 기온은 시정을 악화시키는 것으로 나타났다. 이는 안개가 주로 새벽에 발생한다는 점에서, 주로 새벽의 온도가 낮의 온도보다 낮다는 것을 설명할 수 있다.

상대습도 역시 중요한 피처로 class 1을 제외한 나머지 세 개의 클래스에서 가장 중요한 피처 중 하나로 나타났다. 예상대로 높은 습도는 시정을 악화시키고, 낮은 습도는 시정을 개선하는 경향을 보였다. 이를 통해 안개 발생 시 시정 구간이 1일 경우에는 기온이 주로 영향을 미치고, 시정 구간이 2, 3일 때는 습도가 주로 영향을 미친다는 것을 알 수 있었다.

이러한 결과는 안개 발생을 미리 예측하고 대응하는 데 중요한 정보를 제공한다. 예를 들어, 기온과 습도를 모니터링하여 안개 발생 가능성을 예측하고, 시정 구간을 예측해야 한다. 이를 통해 안개 발생을 더 효과적으로 예측하고, 교통 관리 및 안전 대책을 사전에 준비할 수 있음을 시사한다.

Feature importance에서는 SHAP 분석과 다르게 ws10_deg_sin, ws10_deg_cos, yearly_time_sine 등 값의 크기가 안개 발생과 관련성이 상대적으로 적은 피처들이 높은 중요도로 나타났다. 특히, yearly_time_sine 값은 값의 크기와 안개 발생 정도가 선형적으로 관계있는 것이 아니라, 계절마다 안개 발생 정도가 다르고, 특히 특정 월에 안개 발생이 많기 때문에 중요한 피처로 나타난 것이다. ws10_deg_sin과 ws10_deg_cos는 바람 방향이 안개 형성과 소멸에 영향을 미칠 수 있기 때문에 중요하게 나타난 것으로 해석할 수 있다.

이를 통해, 단순히 feature importance 점수만을 기준으로 모델을 해석하는 것보다, SHAP 값을 활용하여 각 피처가 예측에 어떻게 기여하는지를 면밀히 분석하는 것이 중요함을 알 수 있다. 이는 안개 발생을 예측하고, 교통 관리 및 안전 대책을 마련하는 데 있어 보다 정확하고 신뢰할 수 있는 정보를 제공할 수 있다.



최종 모델의 feature importance

6. 활용 방안 및 기대효과

본 공모전에서 개발한 안개 발생 예측 모델은 다양한 분야에서 실질적인 활용이 가능하다. 시계열 특성을 반영한 예측 모델을 통해 교통 안전성을 향상시키기 위해 도로 교통 관리 기관은 안개 발생 예측 정보를 활용하여 사고 예방 및 교통 흐름 관리에 기여할 수 있고. 항공 운항 안전성을 강화하기 위해 공항 관리 기관은 예측 정보를 기반으로 항공기 운항 일정을 조정하고 안전 조치를 강화할 수 있다. 또한, 기상 서비스 품질을 향상시키기 위해 기상청 등 기상 서비스 제공 기관은 예측 정보를 활용하여 국민에게 정확한 기상 정보를 제공하고 신뢰도를 높일 수 있다. 이러한 활용 방안을 통해 정확한 안개 발생 예측 모델을 개발하는 것은 공공 안전과 서비스 품질을 높이는데 도움이 될 것이다.

*** 개발 환경**

- colab cpu사용

- 라이브러리 버전

NumPy version: 1.25.2/ Pandas version: 2.0.3/ Matplotlib version: 3.7.1/ Seaborn version: 0.13.1/ Scikit-learn version: 1.2.2/ XGBoost version: 2.0.3