기상청

Random Forest를 이용한 기상특성에 따른 안개 발생 진단

참가번호	**240285**	팀명	심정지류

1. 분석 배경 및 목표

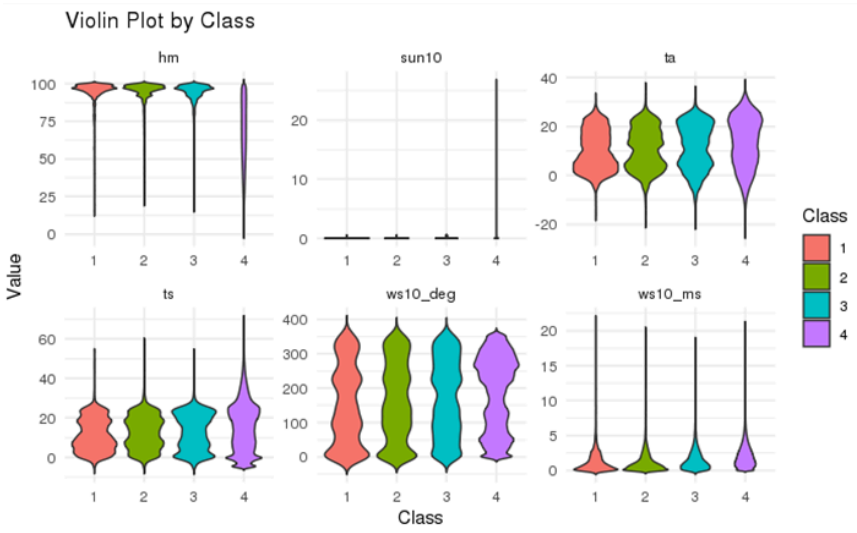
안개는 다양한 분야에서 안전에 영향을 미치는 자연현상으로, 특히 도로 교통사고와 해상 사고 등의 문제를 야기할 수 있다. 이외에도 농업, 생태계와 밀접한 관련이 있어 사회적, 경제적 손실을 초래할 수 있기에 안개 발생을 예측하여 적절한 대응을 하는 것은 중요하다. 그러나 안개 발생에 영향을 미치는 주요 지표인 지면온도, 풍속, 강수 유무외에도 지역과 계절별 특성에 영향을 받기 때문에 예측이 어려운 현상 중 하나이다.
따라서 기상 빅데이터를 활용하여 안개 발생 진단에 대한 정교한 예측 모델이 필요하다.
이를 위해 주어진 5개 지점의 기상 데이터를 활용하여 안개 발생을 예측하는 모델링을 진행하였다.

2. 분석 데이터 정의 및 바이올린 플롯 확인

변수명	내용	변수명	내용
Year	년도	TA	1분 평균 기온 10분 주기, 단위: ℃
Month	월	RE	강수 유무 (0:무강수, 1:강수)
Day	일	HM	1분 평균 상대 습도 10분 주기, 단위: %
Time	시간(0~23)	sun10	1분 일사량 10분 단위 합계, 단위: MJ
Minute	분(10분 단위)	TS	1분 평균 지면온도 10분 주기, 단위: ℃
STN_ID	지점번호	*VIS1	1분 평균 시점 10분 주기, 단위: m
WS10_deg	10분 평균 풍향, 단위: deg	**class**	시정구간
WS10_ms	10분 평균 풍속, 단위: m/s		

<표 1> 사용 변수 정의

변수 확인 과정에서 제공된 데이터를 가공하여 파생 변수를 생성했다. 각 특성에 맞는 모델링을 구축하기 위해서는 기존 변수를 토대로 파생변수를 만들어 모델에 적합시켜야 한다.



<그림 1> 바이올린 플롯

class별로 각 변수들에 대한 바이린 플롯 결과, sun10을 제외한 나머지 변수들에서는 class별로 비슷한 분포를 갖고있다는 것을 알 수 있다. sun10변수의 경우 class가 4인 경우에 대한 분포가 1,2,3에 비해 더 넓게 퍼져있음을 알 수 있으며 이를 통해 sun10과 class 사이에 대해 확인해 볼 필요가 있음을 암시한다.

3. 데이터 전처리 및 변수 추가

3.1 결측값 처리기상_시간선형보간**,** 로지스틱 함수**(**강수 여부**)**

제공 데이터의 평균 기온, 평균 상대 습도, 일사량 합계, 평균 풍향, 평균 풍속, 평균 지면온도, 강수 유무에서 값이 -90이하인 값은 결측치로 고려했으며, 강수 유무를 제외한 나머지값은 시간 선형 보간법을 이용하여 결측치를 채웠다. 또한, 강수 유무에 대한 변수는 로지스틱 회귀 모델을 적합한 값을 이용해 결측치를 채웠다.

3.2 datetime 변수 생성 및 주기성 변수 추가

연도 데이터를 변환해서 연도, 월, 일, 시간, 분을 결합하여 datetime 형식을 지정하고 sin, cos함수를 이용하여 주기성 변수를 새로 생성했다.

3.3 파생변수 추가

3.3.1 위치 변수 **(st)**

stn_id 그룹별로 ta, sun10, ts의 평균을 찾아서 위치변수를 생성했다. 같은 그룹은 같은 위치를 갖도록 st변수를 생성해서 지역별 특성을 반영하고자 했다.

3.3.2 기온과 노점온도 변수 **(ta, tb, ta_tb_diff)**

Magnus-Tetens 공식으로 노점 온도 tb 계산했으며, 기온 ta와 노점온도 tb간의 차이를 계산해 기온과 노점온도 차이 변수를 새로 생성해서 추가했다.

3.3.3 풍속**,** 풍향의 상호작용 변수 **(ws10_new)**

내륙 지역에서는 일반적으로 바람의 영향이 상대적으로 약하지만, 국지적인 풍계의 변화를 파악할 수 있다. 강한 바람은 대기 상태의 불안정성을 나타낼 수 있다. 이를 반영하는

변수를 추가하고자 했다.풍속과 풍향의 상호작용 변수는 안개 발생 메커니즘을 보다 효과적으로 이해하고, 안개 예측 모델의 정확도를 높이는 데 활용될 수 있다.

3.4 이상치 탐지

이상치에 대한 영향을 줄이기 위해 IQR을 사용해 탐지한 후 평균값을 대체했다.

3.5 정규화

min-max 방법으로 평균 풍향, 평균 풍속, 평균 기온 10분 주기, 평균 상대 습도 10분 주기, 일사량 10분 단위 합계, 평균 지면온도 10분 주기의 변수들을 정규화시켰다.

4. 모델링

4.1 모델 구축

안개 발생에 따른 데이터 분포의 불균형을 해결하기 위하여 오버샘플링 기법을 사용하여 target 변수의 분포를 조정한 뒤, 이를 학습데이터와 검증데이터 8:2로 분리하였다. 이후 학습데이터를 분류 모델 RandomForest, XGBoost, ANN(인공신경망)에 적합시킨 후 각 개별 모델의 예측 결과를 결합하는 stacking ensemble 기법을 적용하여 성능을 향상시켰다. 모델 성능 평가는 검증데이터를 활용하였고, F1-score를 기준으로 모델의 정확도를 평가하였다.

4.1.1 XGBoost

XGBoost는 경사 부스팅 기반의 트리 앙상블로, XGBoost는 기존의 gradient boosting 알고리즘을 개선하여 더 빠른 학습 속도와 높은 예측 성능을 보여준다. 또한, 규제 매개변수를 통해 과적합을 효과적으로 제어할 수 있으며, 병렬 처리 기능을 지원하여 대용량 데이터에 대한 처리가 가능하다.

4.1.2 Random Forest

의사결정 트리를 사용하여 회귀 문제를 해결하는 앙상블로, 여러 개의 결정트리를 사용해 평균화하여 과적합을 줄이고 성능을 향상시키고 각 특성의 중요도를 평가할 수 있어 변수 선택에 용이하며 여러 종류의 데이터를 처리할 수 있다.

4.1.3 ANN

다양한 데이터의 패턴을 학습하여 비선형 관계 모델링을 할 수 있으며, 대규모 데이터 처리에 용이하고, 데이터로부터 유용한 특성을 자동으로 추출할 수 있다.

4.1.4 Stacking Ensemble

여러 개의 기본 모델을 조합해 최종 예측을 수행하는 앙상블 기법이다. 기본 모델들이 예측한 결과를 다시 하나의 메타 모델을 통해 결합하여 최종 예측을 수행하는 방식으로 진행된다.

5. 모델 예측 결과

5.1 평가지표

5.1.1 F1-Score

정밀도(Precision)와 재현율(Recall)의 조화 평균을 나타내는 지표이며 정밀도, 재현율은 다음의 식으로 정의된다.

$$\text{Precision (정밀도): } \frac{TP}{(TP+FP)}, \qquad \text{Recall (재현율): } \frac{TP}{(TP+FN)}$$

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

[TP(예측:T, 실제:T), FP(예측:T, 실제:F), FN(예측:F,실제:T), TN(예측:F, 실제: F)]

5.1.2 MAE

회귀 모델의 예측 값과 실제 값 사이의 절대 오차의 평균을 나타낸 값이다.

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\widehat{y_i}\right|$$

5.1.3 Accuracy

모델이 정확하게 예측한 데이터의 비율을 나타낸것으로 분류 모델의 성능을 평가할 때 사용된다.

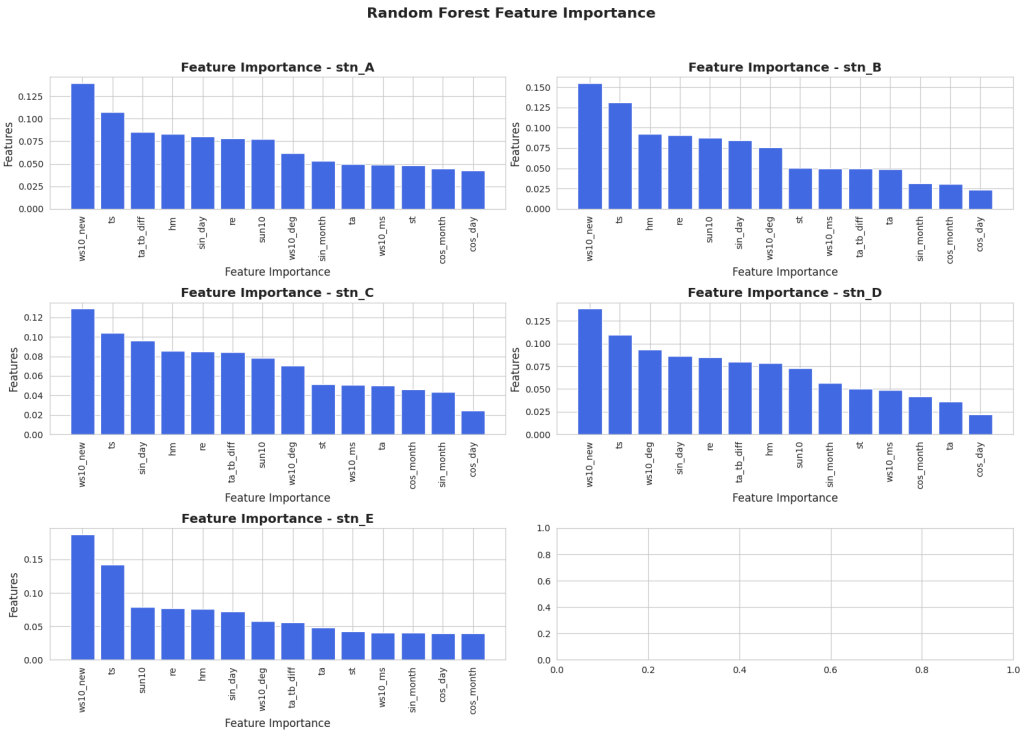
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

5.2 모델별 예측결과

- F1 Score

지역 / 모델	XGBoost	Random Forest	ANN - epochs = 2 - batch_size = 20	Stacking Ensemble
A	0.98493	0.99427	0.97386	0.99335
B	0.95934	0.97939	0.94941	0.97603
C	0.98214	0.99457	0.97389	0.99325
D	0.96456	0.98842	0.94678	0.98463
E	0.98721	0.99515	0.97984	0.99413

<표2> 모델별 예측 성능



<그림2> 변수 중요도

5.3 최종 모델 선정

본 공모전에서는 Random forest 모델을 이용하여 최종 검증을 진행하였다. 지역별 모델에 대한 MAE, Accuracy, F1 score값을 아래 <표3>에 첨부하였으며 F1 score을 기준으로 해당 모델을 선택하게 되었다.

5.4 최종 검증 결과

지역 / 평가 지표	**MAE**	**Accuracy**	**F1 Score**
A	0.00750	0.99427	0.99427
B	0.03198	0.97940	0.97939
C	0.00835	0.99447	0.99457
D	0.01783	0.98842	0.98842
E	0.00665	0.99515	0.99515

<표3> Random Forest 모델로 예측한 지역별 MAE, accuracy, F1 score

6. 한계점

데이터의 불균형으로 인한 오버샘플링 과정에서 데이터의 노이즈도 함께 증폭될 수 있다. 이는 모델의 성능을 저하시킬 수 있으며, 과적합의 위험을 증가시킨다. 또한 오버샘플링으로 생성된 데이터는 기존 데이터의 변형에 불과하므로, 실제로 존재하지 않는 데이터 패턴을 학습할 가능성이 있기에 이는 정확하게 반영하지 못할 수 있다는 한계가 존재한다. 또한 안개 형성은 다수의 기상요소들과 지역의 특성과 밀접하게 관련되어 있다. 다양한 기상 요소와 지역의 특성에 대한 정보가 부족하면 예측 모델이 안개의 형성 조건을 완전하게 파악하지 못하게 됨으로 예측 정확도가 저하 될 수 있기에 이 부분에 있어 아쉬움이 존재하였다.
이외에도 기상데이터는 시간에 따른 수치를 나타낸 시계열 데이터로, 현재의 상태가 이후의 상태에 영향을 주는 유기적인 관계를 가지고 있다고 생각한다. 이를 모델에 반영하여 X_t와 Y_t의 데이터를 $Y_{t+\alpha}$에 적용하여 예측할 수 있었다면 더 좋은 예측 모델을 구축할 수 있었을 것 같다.

7. 활용 방안 및 기대효과

안개 예측은 지역적 특성에 영향을 받기에 예측하는 것이 어렵지만, 교통 사고와 도로 시설, 농업, 생태계 관리 등 다양한 자연 시스템에 영향을 미치기에 중요한 예측 중 하나이다. 안개 형성은 시정 거리와 밀접한 관련이 있으며, 낮은 시정 거리는 교통사고와 도로 시설에 큰 영향을 미친다. 따라서 이를 활용하여 교통 사고 예방에 도움을 줄 수 있을 것이라 기대한다. 이외에도 다양한 기상 및 항공 요소와의 상호작용을 고려하여 안전성, 경제적 이익, 효율적인 자원 관리, 공공 서비스 향상 등 여러 방면에서 안개 발생 예측의 중요성이 점점 더 부각되고 있으며, 정확한 예측을 통해 사회 전반에 걸쳐 긍정적인 효과를 기대한다. 이를

기반으로 지역적 특성과 다양한 기상요소를 고려한 정밀한 예측 모델을 개발하여 활용할 수 있을거라 기대된다.

8. 참고문헌

- Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Ghimire, S., Cerro-Prada, E., Gutiérrez, P. A., Deo, R. C., & Salcedo-Sanz, S. (2022). Machine learning regression and classification methods for fog events prediction. *Atmospheric Research, 272*
- 권철민 (2022). 파이썬 머신러닝 가이드.