기상청

기상위성 자료를 활용한 여름철 자외선 산출 기술 개발

참 가 번 호	220183	팀 명 ※ 반드시 참가신청 시 작성한 팀명	상승기류

□ 분석 개요

○ 연구 배경

- 지구의 성층권 중하부에 위치한 오존층은 광화학 작용을 통해 태양의 유해 자외선을 차단하는 보호막 역할을 수행함
- 최근 성층권 오존은 감소추세이며, 오존이 감소함에 따라 대기 중의 UV-B 흡수가 감소하고 지표에서의 UV-B량이 증가하여 인간에게 피부암, 백내장, 면역체계 악화를 촉진시킴
- 기상청은 전국 7곳에 자외선측정기를 배치해 자외선 복사량을 측정하고 있으며 관측된 자외선 복사량에 특정 파장에 대한 가중치를 곱하고, 구름, 대기 상태, 고도 자료와 결합해 예보함
- 이러한 기상자료를 활용하여 예측된 자외선 지수의 단계에 따른 대응 요령이 존재하며 (그림 1) 의사결정에 도움을 주기 때문에 미래시점에 대한 보다 정확한 자외선 산출 기술의 개발 필요성 증가

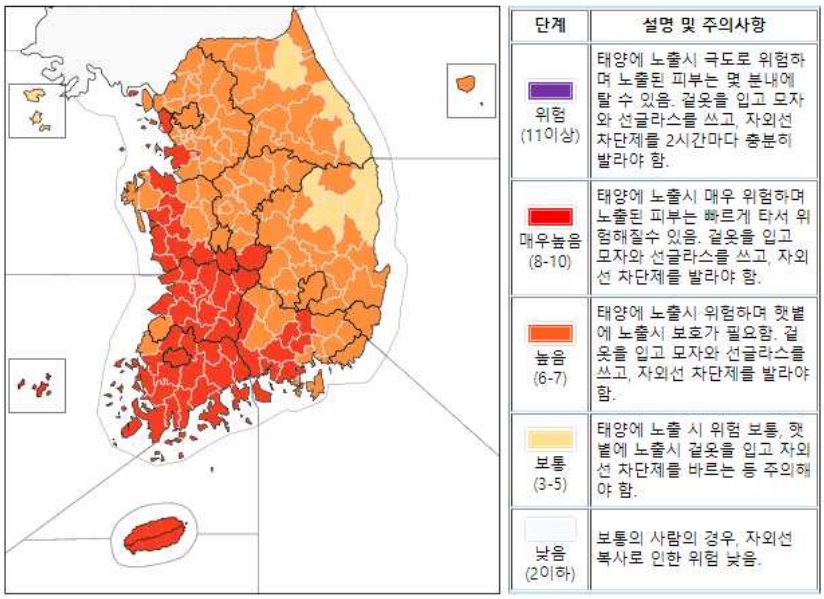


그림 1 지역별 자외선 지수 및 단계별 대응 요령

○ 분석 절차

- 기상위성 자료를 활용한 자외선 산출 기술의 전체적인 절차는 다음과 같음
  1. 데이터 분석 및 전처리(데이터 정의 및 분석, 이상치 및 결측치 처리, 파생변수 생성)
  2. 모델 개발(자외선 예측 모델 개발)
  3. 모델 검증(모델별 비교실험, 최종 모델 성능 검증)

□ 데이터 분석 및 전처리

○ 데이터 정의

- 기상청의 두 번째 기상위성인 천리안위성 2A호의 관측 데이터(그림 2)를 사용하며, 총 16개의 채널이 존재(가시채널 4개, 적외채널 12개)

그림 2 천리안위성 2A호의 16개 채널 관측데이터

표 1 변수 정의 테이블

변수명	정의
yyyymmdd	YearMonthDay
hhnn	HourMinute
STN	관측소지점
Lon	경도
Lat	위도
UV	자외선
Band1	파랑 가시밴드
Band2	초록 가시밴드
Band3	빨강 가시밴드
Band4	식생 가시밴드
Band5	권운 밴드
Band6	눈/얼음 가시밴드
Band7	야간안개/하층운 밴드
Band8	상층 수증기 밴드

변수명	정의
Band9	중층 수증기 밴드
Band10	하층 수증기 밴드
Band11	구름상 밴드
Band12	오존 밴드
Band13	대기창 밴드
Band14	깨끗한 대기창 밴드
Band15	오염된 대기창 밴드
Band16	이산화탄소(CO2)밴드
SolarZA	태양천정각
SateZA	위성천정각
ESR	대기외일사량
Height	관측고도
LandType	지면타입

- 입력변수: 자외선 변수를 제외한 데이터 관측 시간을 나타내는 두 변수('yyyymmdd', 'hhnn'), 관측 지점에 대한 정보를 나타내는 세 변수('STN', 'Lon', 'Lat'), 기상위성에서 관측한 16개의 채널 변수('Band1' ~ 'Band16') 및 태양천정각, 위성천정각, 대기외일사량, 관측고도, 지면타입 변수를 포함하여 총 26개의 변수가 존재

- 목적변수: 자외선 변수('UV')

○ 데이터 분석

- 학습 데이터의 경우 2020.01.01. ~ 2021.12.31. 까지 총 2년 치의 데이터가 존재하며 전처리 시 변수 형태 균일화, 결측값 대체 및 이상치 처리, 범주형 데이터의 0~1 범위 수치화, 관측 시간에 따른 cyclical encoding, min-max 정규화 등을 수행
- 데이터 로드 시 변수명은 '{관측시점_uv}.변수명' 으로 이루어짐. 따라서 하나의 데이터 프레임으로 만들기 위해 {관측시점_uv}를 제거 후 '변수명'만으로 이루어진 하나의 데이터 프레임으로 병합
- 'yyyymmdd', 'hhnn'변수를 활용하여 시간을 나타내는 'Date_Time' 변수를 생성 후 'stn', 'Date_Time' 변수를 활용하여 지점별 시간 순으로 정렬
- 병합된 데이터의 결측행을 구했을 때, 'uv'변수가 53207행, 'Band1' ~ 'Band16' 16개의 변수에서 각각 18060행의 결측이 존재함을 확인
- 결측값에 대한 보간 처리 시 여러 방법 중 'time' 방식을 선택, 'linear' 또는 'poly' 방식으로 보간 시 데이터의 주기적 특성이 반영되지 않는 문제 발생

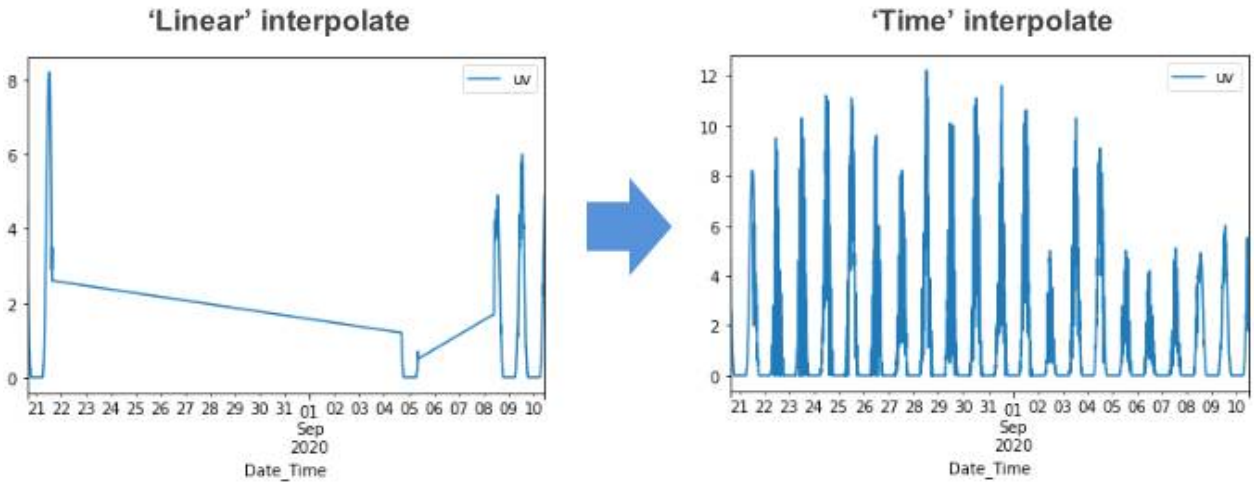


그림 3 13번 지점의 결측지점 보간 방식에 따른 보간 결과

- 그림 3은 13번 지점의 보간 방식에 따른 보간 결과의 차이를 보여주는 그림으로, 데이터의 인덱스를 반영하여 보간을 수행할 때 주기적 특성이 보존되는 결과를 얻음
- 따라서, 주기적 특성이 보존되는 보간을 지점마다 수행하여 15개 모든 지점에 대해서 각각 적용
- 보간 시, 'uv' 변수의 값에서 직전 시점과의 차이 값을 계산한 'diff' 변수를 생성 후 일정 값 (3 표준편차) 이상으로 급격히 변하는 구간은 nan처리 후 재보간 수행

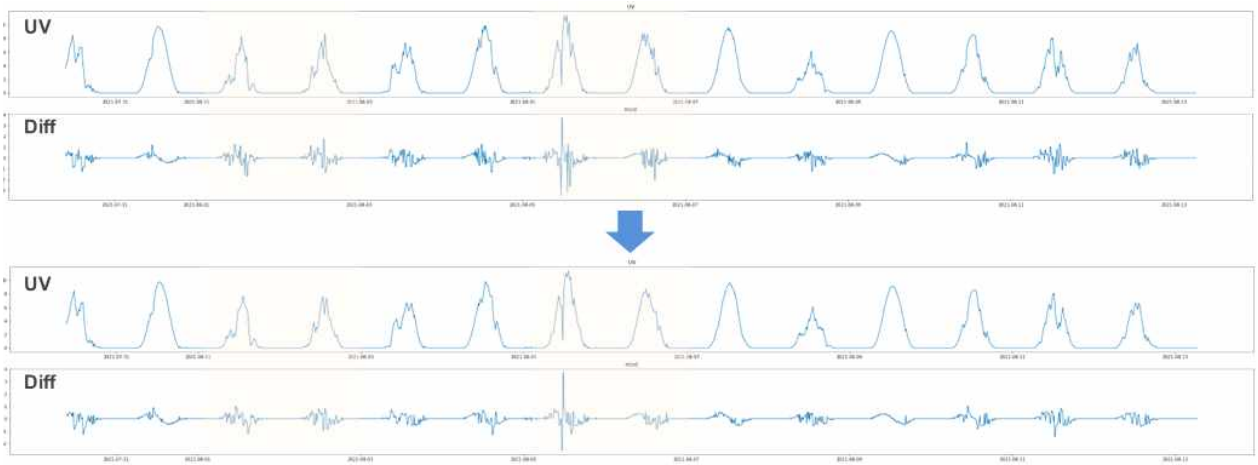


그림 4 165번 지점에서 총 5번의 보간을 반복 수행 후 변화된 'uv', 'diff' 값

- 그림 4는 165번 지점의 보간을 5회 반복 수행한 경우 값의 변화를 나타내는 그림으로 보간을 한 번 수행했을 때보다 여러번 반복 수행한 경우 'UV', 'Diff' 변수의 분포가 완만해지는 것을 확인할 수 있음
- 보간을 통해 이상치 및 결측치를 처리한 후, 15개의 범주를 가지는 'sateza', 'height' 변수 및 4개의 값을 가지는 'landtype' 변수는 더미화를 수행하였으며, 'stn' 변수의 경우 0~1 사이의 값을 가지도록 수치화하였음
- 데이터의 시간적 정보를 반영하기 위해 'Date_time' 변수에 대해 sine 및 cosine 함수를 이용한 cyclical encoding을 수행

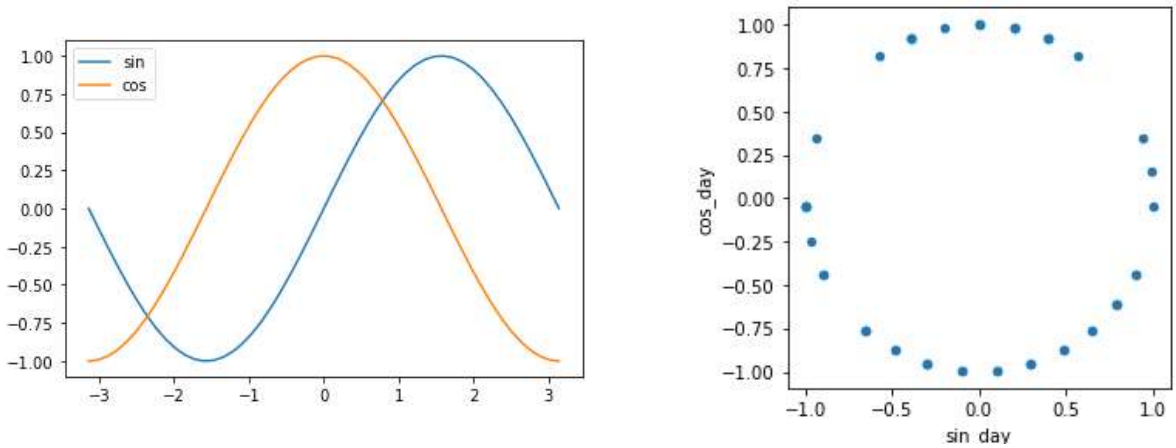


그림 5 (좌) sine, consine 함수, (우) sine, cosine 변환을 통한 일(day) 변환

- $x_{\sin} = \sin(\frac{2\times\pi\times x}{\max(x)}), \quad x_{\cos} = \cos(\frac{2\times\pi\times x}{\max(x)})$
- 그림 5는 sine 및 cosine 변환을 나타내며. cyclical encoding을 통해 모델이 잘 학습할 수 있는 변수로 변환. x는 변수를 의미하며, 시간(hour)을 기준으로 주기는 24시 이므로 $\max(x)$는 24를 의미하며, 일(day)을 기준으로 주기는 31일이며 $\max(x)$는 31을 의미함

□ 자외선 산출 모델 개발

○ 모델 개발

- 전처리된 데이터에 시간정보를 반영한 변수를 생성하였기 때문에, 입력데이터가 시간 순으로 들어가야 하는 시계열 학습 방식이 아닌 tabular data로써 다양한 모델 적용 가능하며 tabular data에 우수한 성능을 보이는 머신러닝 및 딥러닝 모델 적용
- 시도 모델: 머신러닝 모델(LightGBM) 딥러닝 모델(DNN, TabNet, LSTM), 앙상블 모델(Soft voting ensemble)
- 후보 모델: LightGBM, DNN, TabNet
- 모델 제외 사유: LSTM의 경우 window size 이용하여 데이터를 구성하는데, window size 이후 시점을 예측해야하기 때문에 학습 시 처음의 window size 만큼은 예측이 불가

○ 실험 설계

- 평가 지표: RMSE(Root Mean Square Error)를 사용하였으며, 잔차의 제곱합을 산술평균한 값의 제곱근으로 관측값들 간의 상호간 편차를 의미함. 표준편차를 일반화시킨 척도로 실제값과 추정값의 차이를 알려줌(RMSE가 1일 때, 자외선 지수가 대략 1정도 차이가 난다는 것을 의미)
- $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(F_i - O_i)^2}$ $(F_i : 관측값, O_i : 예측값)$
- 학습 및 검증데이터 분리: 5fold cross validation 수행, tabular data 관점으로 데이터 전처리를 진행했기 때문에 시간 순서에 관계없이 데이터 분리 가능
- 후보 모델들의 하이퍼 파라미터는 wandb 라이브러리를 사용하여 최적의 값을 찾아가며 하이퍼 파라미터 서치를 하였으며, soft voting ensemble 및 data augmentation 기법을 활용하여 성능 향상을 시도

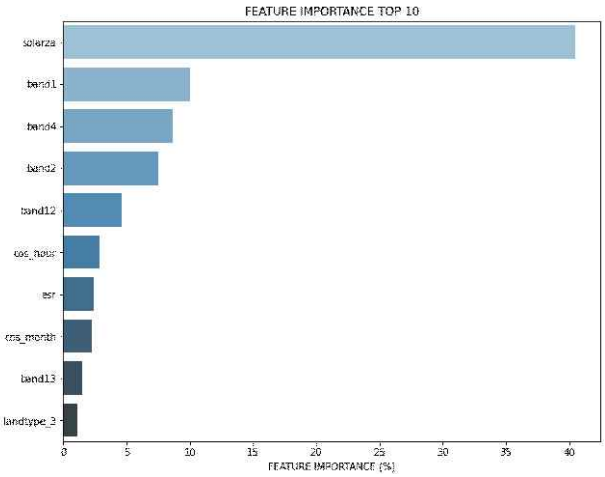
○ 실험 결과 및 결론

Model	Post-processing O	Post-processing X
	RMSE(↓)	RMSE(↓)
LightGBM	0.629798	**0.629760**
TabNet	**0.613479**	0.613494
DNN	0.604392	**0.604293** ✓

그림 6 모델별 성능 비교 결과 표

- 그림 6의 post-processing은 예측 결과의 후처리를 의미하며, 학습 데이터에서의 0을 제외한 최솟값보다 평가 데이터에서 예측한 값이 작은 경우 해당 'UV'값에 대해 0처리를 해준 것
- 실험 결과 DNN 모델이 LightGBM과 TabNet보다 우수한 결과를 보임을 알 수 있고, 후처리의 여부가 유의미한 성능 변화를 주지 않는 것을 볼 수 있음
- 따라서, 후처리를 하지 않은 DNN 모델을 최종 모델로 선정하였으며, 해당 최종 모델로 중요 변수 순위화 및 중요 변수 선택에 따른 성능 변화 등의 추가 실험을 진행

- eXplainable AI(XAI) 기법을 통한 변수 중요도 계산



Model	Post-processing O	Post-processing X
	RMSE(↓)	RMSE(↓)
TOP 30	0.609335	**0.609270**
TOP 40	0.599597	**0.599517**
TOP 50	0.599045	**0.596673**
Ensemble	0.592531	**0.592461** ✓

참가번호 **220183** 의 RMSE 값은 **0.592461** 순위는 **3**위 입니다.

그림 7 (좌) SHAP을 통한 상위 10개 중요 변수, (우) 변수 개수에 따른 성능 비교 결과 표

- XAI 기법 중 SHAP 방법론을 통해 최종 선정 모델의 변수 중요도를 계산하였으며, 상위 10개 변수가 차지하는 중요도는 약 81%를 넘어섬(그림 7의 좌측)
- 상위 10개의 변수 중요도는 다음과 같음. 'solarza': 40.47%, 'band1': 9.98%, 'band4': 8.65%, 'band2': 7.48%, 'band12': 4.62%, 'cos_hour': 2.83%, 'esr': 2.41%, 'cos_month': 2.23%, 'band13': 1.47%, 'landtype_3': 1.14%. 도출된 결과에 따라 태양 천정각을 의미하는 'solarza' 변수가 가장 중요한 요소로 파악됨
- 중요 변수만을 선택하여 학습 시, 상위 50개의 변수를 선택하여 학습한 경우 RMSE는 0.596673으로 모든 변수를 사용한 결과보다 RMSE 성능이 0.00762 만큼 좋은 것을 볼 수 있음(그림 7의 우측)
- 결과가 가장 좋게 도출된 상위 50개의 변수만 사용하여 하이퍼 파라미터에 변화를 주면서 추가 학습을 진행하였고, 학습된 모델들의 앙상블을 통해 RMSE가 0.592461로 앙상블을 하지 않았을 때보다 성능이 좋은 모델을 도출함
- 본 연구의 결론은 다음과 같음. 후보 모델 중 DNN 모델이 가장 우수한 결과를 얻을 수 있었으며 후처리의 여부가 유의미한 성능 차이를 주지 않는 것을 확인 할 수 있음
- 또한, 최종 선택 모델에서 XAI 방법론을 통해 가장 중요한 변수를 구했고, 그 결과 태양 천정각을 의미하는 'solarza' 변수가 40.47%의 중요도를 가지면서 가장 중요한 변수로 파악됨
- 모든 변수를 사용할 때보다 변수 중요도 기준 상위 50개의 변수를 선택하여 학습할 때 더욱 좋은 모델을 구축할 수 있으며, 단일 모델 보다 앙상블 시도 시 보다 좋은 모델 구축 가능

□ 기대효과

- 본 연구의 경우, 위성 영상데이터만을 이용하여 자외선 산출 모델을 개발하였지만, 추후 예보 데이터 및 기상관측 데이터와 결합하여 보다 더욱 정확한 지역별 자외선 지수 산출 모델을 개발할 수 있음
- 보다 적은 변수를 사용하여 높은 정확성을 가지는 자외선 지수 산출 모델의 개발이 가능해지며 현재 운영 중인 5단계의 자외선 지수에서 보다 여러 단계로 세분화 가능
- 자외선 지수 산출 모델의 예측 결과를 바탕으로 단계에 따른 대응 요령 등 주의보/경보 시스템 구축화가 가능하며 자외선과 관련된 국민보건문제를 최소화 가능

□ 참고 문헌

1. Kim, Yu-Geun, Hwa-Un Lee, and Yun-Seop Mun. "한반도 지역의 유해자외선 (UV-B) 강도 및 자외선 지수 예측 시스템에 관한 연구." Proceedings of the Korea Air Pollution Research Association Conference. Korean Society for Atmospheric Environment, 1998.

2. 강나경. 자외선 (UV-B) 노출에 따른 인체위해도 예측모델의 고찰. Diss. 연세대학교 보건대학원, 1996.

3. 자외선 지수 예보정보(웨더아이) https://c.weatheri.co.kr/forecast/forecast09.php

4. 생활기상지수-자외선지수(기상청 날씨누리)
   https://www.weather.go.kr/w/theme/daily-life/life-weather-index.do

5. 서동준. (2022년 06월 05일). 자외선 지수가 알려주는 지구환경. 동아사이언스
   https://www.dongascience.com/news.php?idx=54727

6. 김민주. (2019년 10월 30일). 자외선차단 수치를 더 정확히 예측하는 머신러닝 모델
   https://www.thekbs.co.kr/news/articleView.html?idxno=1610

□ 활용데이터 목록

1. 기상위성 자료 (기상청)
2. 자외선지수 자료 (기상청)