강할 땐 무섭고 약할 땐 반가운 자외선지수 예측

참 가 번 호	220090	팀 명	SUV

1. 배경 및 분석목적

최근 급격한 기온 상승으로 인해 심각한 기온 변화 문제에 대한 관심이 높아지고 있다. 이러한 관심을 바탕으로 다양한 뉴스를 찾아보던 중 예년에 비해 올해 열대야 시작일이 비교되지 않을 정도로 빨라졌다는 사실을 알 수 있었다.[1] 열대야 시작일이 빨라진다는 것은 여름의 시작이 빨라진다는 말과 같다.

우리나라의 경우 여름철 자외선지수가 연중 가장 높은 수치를 보이기에 여름이 빨라짐에 따라 자외선지수가 높은 수치를 기록하는 일수가 더 길어질 것이라는 생각이 들었다. 이러한 호기심을 가지고 있던 중, '날씨 빅데이터 콘테스트' 과제 중 '여름철 자외선지수 예측 모델 개발'을 확인하였고, 참가하게 되었다. 데이터 분석을 위해 자외선에 대하여 조사해 보았다. 자외선이란, 가시광선보다 짧은 파장으로 눈에 보이지 않는 빛이며, 자외선의 파장의 길이에 따라 UV-A~C로 나눌 수 있다. 자외선의 노출은 비타민-D 합성이라는 유익한 점도 있는 반면, 탈모, 피부암, 백내장 등의 인체에 유해한 영향을 주기도 한다. 추가적인 자외선과 관련된 정보 수집 중, "자외선지수는 계절과 그날 날씨에 따라 오르고 내리기도 하지만, 지난 수 십년간 전 지구적으로 지속 증가한 것으로 나타난다[2]."라는 것을 확인하였다. 이러한 배경을 토대로, 자외선이 가장 높다고 예측되는 여름철의 자외선 값을 예측하는 모델을 생성하여, 자외선 값에 대한 정보를 제공하고, 최종적으로 현재 흐름에 따른 자외선 값을 파악하고자 한다.

2. 데이터 전처리 및 탐색

2-1 데이터 전처리

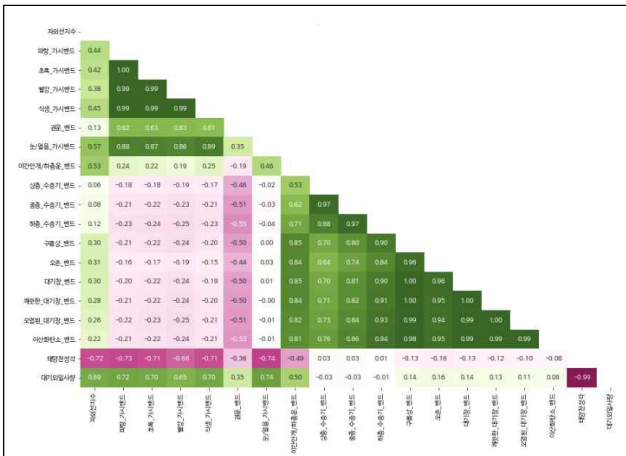
데이터 통계량을 확인한 결과 -999 값과 음수 값이 존재한 것을 확인할 수 있었다. 먼저, -999 값은 일반적으로 누락된 값으로 다른 값들과 구분하기 위해 전혀 연관되지 않은 값으로 볼 수 있었다. 따라서 -999 값은 일괄 결측치로 처리한 후 결측치 보간을 진행할 것이다. 결측치 확인 결과 자외선지수 53,207개, 그 외의 밴드들에는 18,060개가 있음을 확인하였다.

다음으로, 가시채널과 근적외채널의 밴드에 음수값이 존재하였다. 멘토링 결과 두 채널의 특성상 야간에는 측정이 불가능하고 음수값이 존재할 수 없다는 것으로 확인되었다. 이를 주간 야간으로 구분하여 전처리하기 위해 변수 중 태양천정각을 이용했으며 태양천정각의 90도를 기준으로 아래로는 주간, 위로는 야간임을 확인했다. 이를 통해 가시채널과 근적외채널 중 야간의 값과 주간의 음수값을 0으로 처리했다. 오입력된 값들을 삭제하지 않고 0으로 처리한 이유는 데이터 삭제 없이 최대한의 데이터를 분석하기 위함이다. 다양한 결측치 보간법 중 하나를 선택하기 위해 회귀모델 중 엘라스틱넷을 통해 값의 결과가 가장 좋은 interpolate 결측치 보간법을 이용하기로 하였다. interpolate란 값에 선형으로 비례하는 방식으로 결측값을 보간하는 함수이다. 결측치를 보간하는 과정에서 전체 데이터와 자외선 결측치가 포함된 행을 지운 데이터의 결과를 비교하기 위해 두 데이터의 결측치를 보간한 후 회귀모델인 엘라스틱넷을 통해 더 유의미한 결과인 전체 데이터를 사용하기로 하였다.

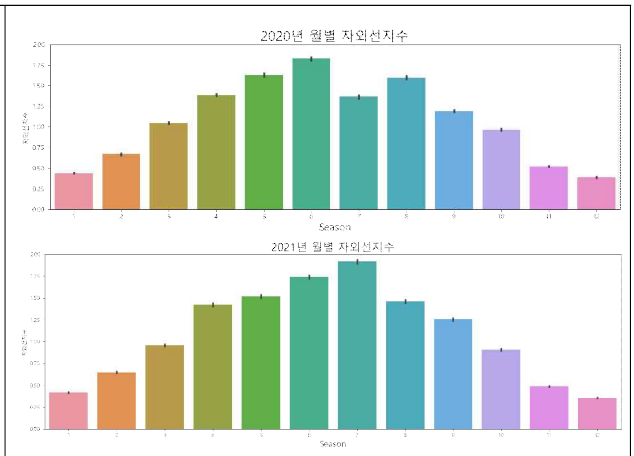
분석에 활용한 데이터의 범위의 경우, 검증데이터의 기간이 6월인 것을 감안해 3월부터 7월까지의 데이터를 사용할 예정이다. 변수를 줄이기 위해 5개의 채널별 밴드들의 통계량 수치와 상관관계를 확인한 후 평균을 이용하여 하나로 묶어주었다. 또한 단파적외채널, 수증기채널, 적외채널의 단위가 K(Kelvin)으로 같으므로 세 개의 채널을 하나로 다시 묶어주었다. 변수 중요도를 확인한 결과 채널1과 채널2는 높은값의 결과가 나왔기 때문에 추가 전처리하지 않았다. 현재 변수 중 태양천정각을 기준으로 자외선지

수가 달라질 것으로 예상하여 0~45 이하, 45~90 이하, 90~135 이하, 135~180 이하로 나눈 후 새로운 컬럼 'time'을 생성하였다. 변수 중 관측소지점, time은 범주형 데이터이므로 숫자형으로 처리하기 위해 원핫인코딩을 진행하였다. 지면타입의 경우, 하나의 관측소지점에 하나의 지면타입을 사용하기 때문에 삭제해 주었다.

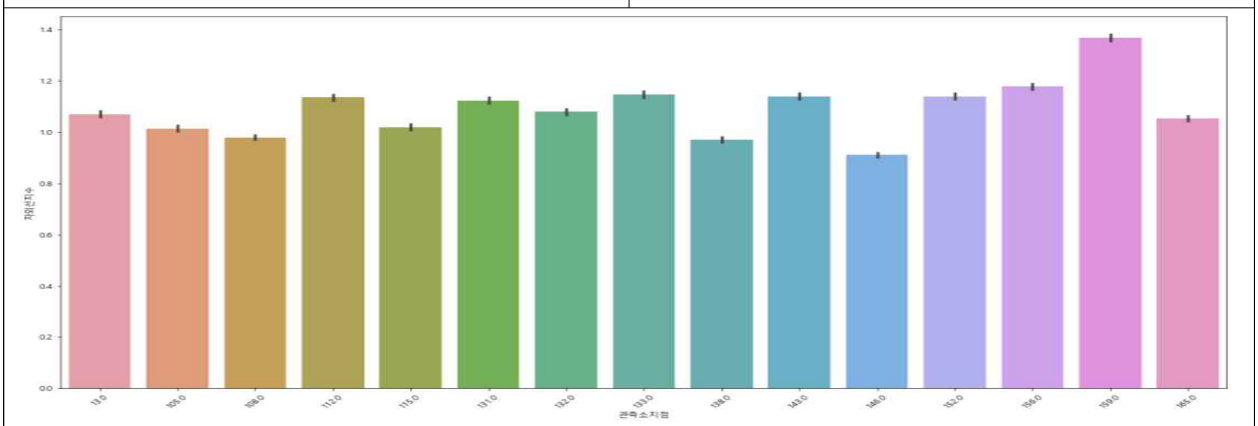
2-2 데이터 탐색



[그림 1] 상관관계가 높은 변수들의 히트맵



[그림 2] 연도별 월별 자외선지수



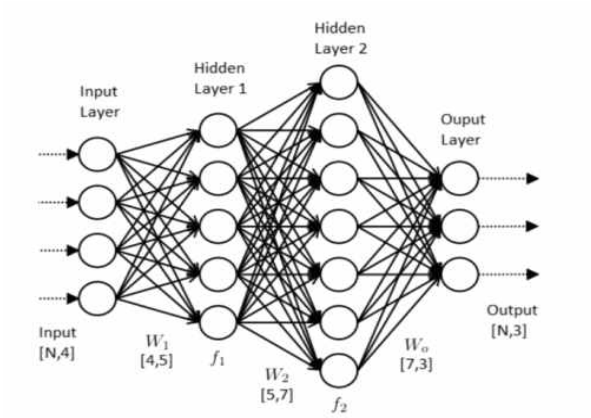
[그림 3] 관측소지점별 자외선지수

- 수치형 변수별 상관계수를 확인해보았을 때, 가시채널의 밴드들간의 상관계수가 높게 나타났다. 가시채널과 근적외채널은 태양천정각에서 음의 상관관계, 대기외일사량에서는 양의 상관관계가 높게 나타난다. [그림 1]
- 2020년과 2021년의 월별 자외선지수를 나타낸 막대그래프이며 이를 확인한 결과, 두 해 모두 정규분포를 따르고 있음을 보인다. 2020년은 6월, 2021년은 7월에서 가장 높은 자외선지수가 나타났다. 2020년 7월은 한여름이지만 자외선지수가 낮게 나타나는 것을 확인할 수 있었다. 이는 7월에 북태평양고기압이 북쪽으로 확장이 지연되는 가운데, 북쪽의 찬 공기와 만나 활성화된 정체전선이 우리나라를 오르내리며 장마철이 길게 이어져 전국 강수량이 평년보다 많았으며 그로 인해 이상저온이 발생하여 자외선이 평년보다 낮았던 것으로 보여진다. [그림 2] [3]
- 관측소지점별 자외선지수를 막대그래프로 확인한 결과, 관측소지점 159(부산)이 가장 높고, 146(전주)이 가장 낮다. [그림 3]

3. 모델 설명 및 성능평가

3-1 모델 설명

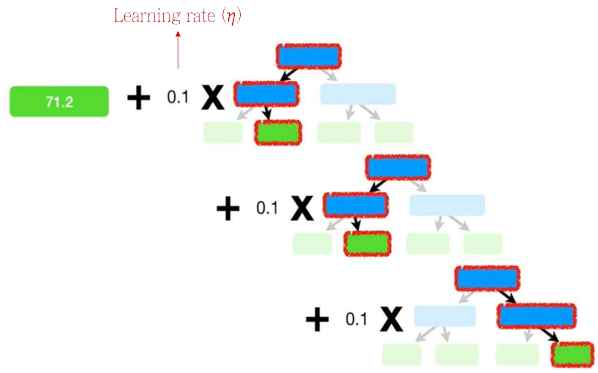
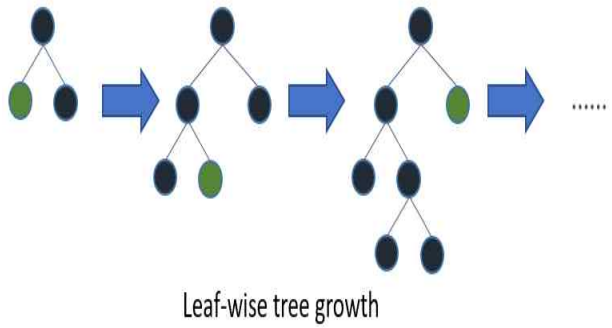
(1) Multy Layer Perceptron Regression(MLP)

Hidden Layer form of MLP	Relu 함수 계산 식[8]
	$ReLU(x) = \max(0, x)$

[그림 4] MLP 모델의 작동원리 및 회귀 활성화 함수

- MLP는 입력층과 출력층 사이에 하나 이상의 은닉층이 존재하는 인공신경망으로 아래의 형태를 가진다. MLP의 입력층은 다수의 입력 데이터를 받고, 출력층은 데이터의 출력을 담당한다. 은닉층은 입력층과 출력층 사이에서 두 층을 연결시킨다. 은닉층의 층수와 각 층의 노드 개수를 설정하여 모델을 구성한다[4].
- ReLU 함수는 0보다 큰 입력값의 경우 그대로 출력하며, 0 이하의 값은 다음 층에 전달을 하지 않는 방식이다. 이후 출력된 값과 실제값의 차이를 최소화하는 방향으로 학습이 반복되는 오차 역전파 알고리즘을 통해 학습이 수행된다.
- 경사 하강법을 통하여 출력값과 입력값의 차이를 줄여주는 방향으로 은닉층과 출력층의 가중치를 재설정 해 나간다. 본 데이터는 1,000,000개 이상의 대규모 데이터이므로, 분석을 진행할 때, solver를 'adam'으로 설정하였다. 'adam' 옵티마이저는 'Momentum' 옵티마이저와 'RMSprop' 옵티마이저의 방식을 섞어놓은 알고리즘으로써, 대규모 데이터의 학습 시 비교적 빠른 속도와 모델의 정확성을 따라갈 수 있다는 장점이 있다.

(2) Gradient Boosting Decision Tree Model & Light Gradient Boosting Model

Gradient Boosting Decision Tree Model[9]	Light Gradient Boosting Model[10]
	

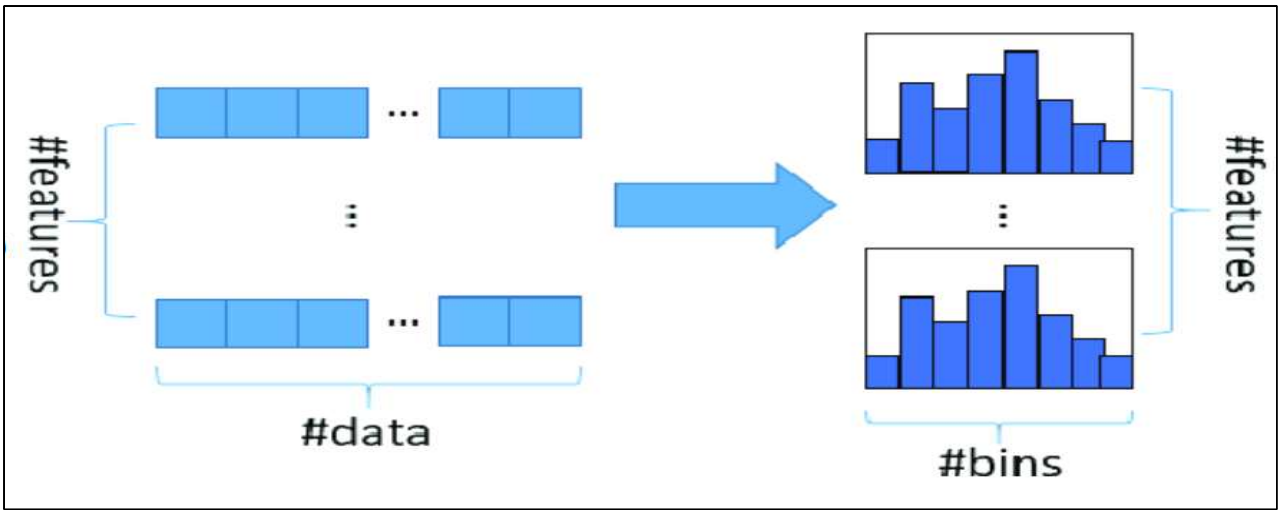
[그림 5] GBT 모델과 LightGB 모델의 작동원리

- Gradient Boosting(GBT) 모델은 트리를 생성하여 각 output의 평균이 leaf가 가지는 값이 되며, 이후 잔차를 계산한다. 이 과정을 반복하여 이전 트리의 잔차를 점차 줄여나가는 기법이다. GBT 모델은 트리의 깊이를 뜻하는 max_depth의 값이 3~6인 얕은 의사결정 나무일 때 높은 예측력을 보이는 특징

이 있다[5].

- GBT 모델의 경우, 트리 분할 방식이 균형 트리 분할 (level–wise) 방식을 채택하기 때문에 모델의 훈련 시간이 오래 걸리며, 메모리 소비가 크다는 단점이 존재한다[6]. 이를 극복하여 만들어진 알고리즘이 LightGBM이다. LightGBM은 **GOSS(Gradient Based One Side Sampling)**와 **EFB(Exclusive Feature Bundling)** GBT 모델의 가장 큰 단점인 데이터 처리시간을 보완한 알고리즘으로[7], 리프 중심 트리 분할(Leaf–wise) 방식을 통해 max delta loss(최대 손실 값)를 가지는 노드를 지속적으로 분할하기 때문에 비대칭 트리가 만들어지게 된다.
- 이러한 LightGBM 모델은 빠른 학습 및 예측 수행 시간, 더 적은 메모리 사용량, 높은 정확성, 병렬 계산 및 분할, GPU 학습지원, 대규모 데이터 처리 유용 등의 특징을 가지고 있다. 단, LightGBM을 사용하기 위해서는 데이터의 개수가 10,000개 이상이어야 하는데, 만약 데이터의 수가 적을 경우, 과적합의 문제가 발생할 가능성이 높다.

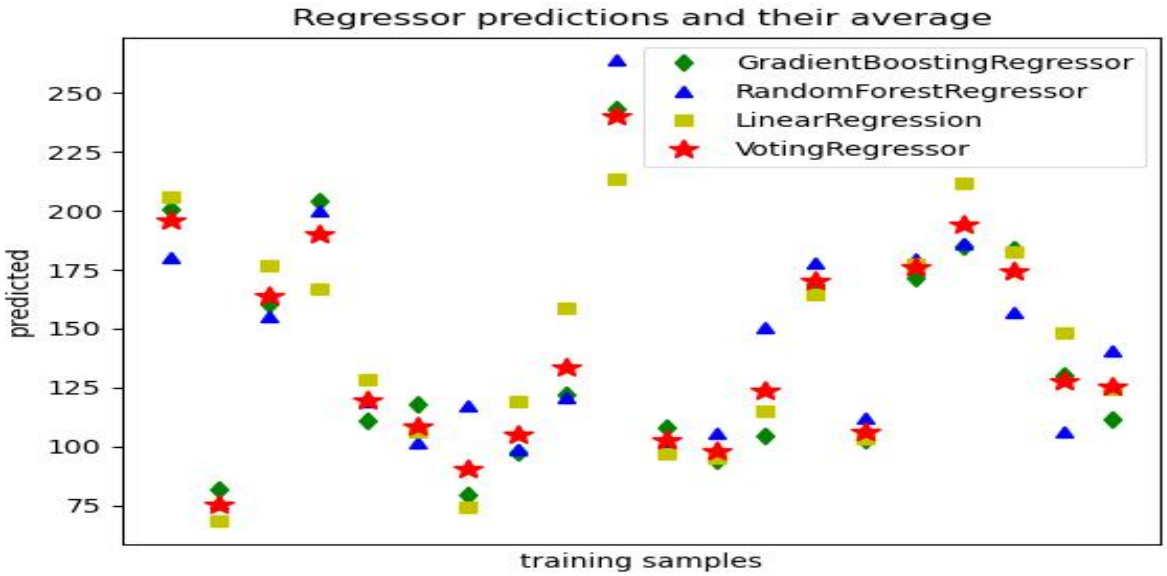
(3) Histogram-Based Gradient Boosting Model



[그림 6] Histogram based Gradient Boosting 모델의 작동 원리[11]

- Histogram-Based Gradient Boosting(HGBT) 모델은 LightGBM과 매우 유사한 모델로, LightGBM의 특징과 더불어 결측치에 대한 imputer가 필요하지 않다는 추가적인 특징을 가지고 있다. 다만, LightGBM과 달리 전처리 과정에서 사용하는 binning을 의사 결정 트리 알고리즘에도 적용하여 알고리즘의 속도를 높인다는 것에 차이가 있다. 이때, 모델 스스로 최적의 bins를 찾아내 고려할 분할점의 수를 줄이며(일반적으로 최대 bins는 256), 정렬된 연속 값에 의존하지 않은 데이터 구조를 활용할 수 있다.
- HGBT 모델의 parameter 중 max_iter는 GBT 모델의 parameter 중 n_estimator와 같은 역할을 한다. 예측을 위한 훈련 데이터의 양이 많을 경우, 데이터의 크기를 줄이기 위한 일반적인 방법은 다운 샘플링 기법을 사용한다[6]. 하지만 다운 샘플링을 할 경우, 데이터를 손실한다는 단점이 존재하는데, HGBT 모델의 경우, 모델 학습의 속도가 빠르기 때문에 다운 샘플링이 필요 없고, 그 결과 데이터 손실을 최소화한다는 큰 장점이 있다.

(4) Voting Regressor



[그림 7] 여러 모델들의 예측 값

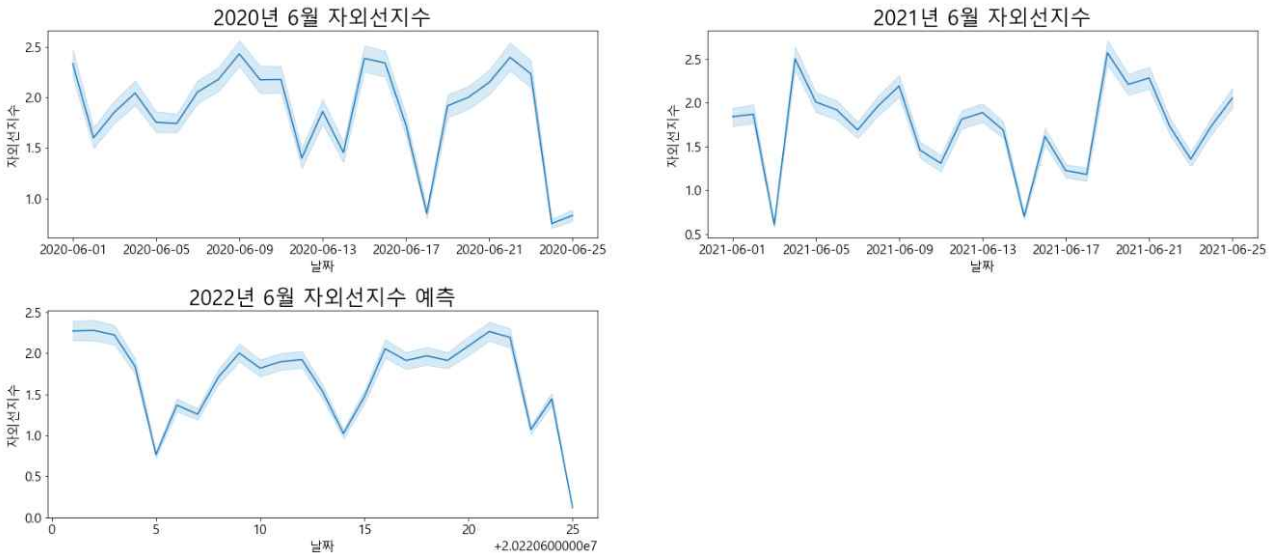
- Voting 기법의 경우, 한 모델에서 다른 sample 조합을 사용하는 bagging 기법과 달리, 서로 다른 모델 간의 조합을 통해 평균 예측값을 반환하는 기법이다. 이 기법의 경우 한 모델에 대한 약점을 극복할 수 있는 장점이 있다.

3-2 성능 평가 결과

모델	MLP	GBT	HGBT	**Voting**
RMSE	0.647	0.677	0.675	**0.631**

- 모델 검증 결과, GBT 모델과 HGBT 모델을 Voting 한 모델의 RMSE가 0.631로 가장 작게 나타났다.

4. 활용방안 및 개선방향



[그림 8] 분석 예측 결과

분석 모델(GBT, HistoGBT Voting 모델) 예측 결과 2020년, 2021년에 비해 모델이 예측한 2022년 자외선지수가 상대적으로 변동이 적을 것이며, 자외선지수 수치 또한 비교적 높을 것으로 예측했다.

이러한 분석결과를 바탕으로 자외선예측 모델을 이용한 활용방안은 다음과 같다.

우선 안구건강과 피부건강에 악영향을 미치는 자외선지수 예측을 통해 많은 병들을 예방할 수 있다. 또한 높은 자외선지수가 초래하는 각종 병들의 위험도를 표시하여 어린이, 노약자 등 건강에 많은 관심을 필요로하는 연령대가 있는 가구에 자외선 경보를 발송한다면 건강관리에 도움을 줄 수 있을 것이다.

이와 반대로 비타민D, 살균작용 등 자외선이 주는 이로운 영향을 극대화할 수 있도록 자외선지수 예측 결과에 따른 야외활동 지수를 같이 표시한다면 자외선을 보다 이롭게 이용할 수 있을 것으로 기대된다.

해당 분석의 한계점은 자외선지수에 중점을 둔 분석이기에 활용방안이 다양하지 않다는 것에 있다. 따라서 분석결과와 연계할 수 있는 다양한 활용방안을 제시하고자 한다.

(1) 지역별 자외선지수 예측

시도별 위도, 경도 데이터와 기후 및 식생 데이터를 통해 자외선 예측을 진행한다면 보다 세밀한 자외선 경보 시스템을 개발할 수 있다. 이렇게 개발된 시스템을 바탕으로 자외선지수가 높은 지역과 낮은 지역의 특징을 파악하여 정확한 자외선 지수 예측 모델 구축이 가능해질 것으로 기대된다.

(2) 오존층 파괴 및 오염가스 배출 관리

2021년 유럽연합 코페르니쿠스 대기 모니터링 서비스팀은 현재 남극 오존층 구멍이 어느 때보다도 더 크다는 관측 결과를 발표한 바 있다. (사이언스타임즈, 2021)

전세계 국가들의 노력에 따라 오존층이 다시 회복추세로 들어갔다고 봤으나 2021년 다시 오존층이 파괴되고 있으며, 이에따라 국내 오염가스 배출 관리가 시급한 상황이다. 따라서 오존층 파괴에 따른 자외선 지수의 증가를 해당 분석 결과와 연계하여 오염가스 배출 현황과 오존층 위성사진을 같이 제공한다면 오존층 파괴 문제의 심각성을 제시함과 더불어 경각심을 심어줄 수 있을 것이라 기대된다.

참고 문헌

[1] https://news.sbs.co.kr/news/endPage.do?news_id=N1006816491
[2] https://www.dongascience.com/news.php?idx=54727
[3] https://www.kma.go.kr/download_02/ellinonewsletter_2020_07.pdf : 2020년 7월 기후동향
[4] 홍석경,안재훈(2021) 다층퍼셉트론과 합성곱 신경망에 기반한 지진 지반응답해석. 한국방재학회 한국방재학회논문집,231-238
[5] B. Ilyasov, E. Makarova, V. Martynov, E. Zakieva, E. Gabdullina and M. Yusupov, "Application of Gradient Boosting Algorithm for Predicting Equipment Failures," 2022 VI International Conference on Information Technologies in Engineering Education (Inforino), 2022, pp. 1-5, doi: 10.1109/Inforino53888.2022.9783011.
[6] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree."Advances in neural information processing systems 30 (2017).
[7] 장승일, 곽근창.(2019).XGBoost와 LightGBM을 이용한 안전 운전자 예측 성능 비교.한국정보기술학회 종합학술발표논문집,(),360-362.

사진 출처

[8] https://yhyun225.tistory.com/21
[9] https://tyami.github.io/machine%20learning/ensemble-4-boosting-gradient-boosting-regression/ (GBT 모델 사진)
[10] https://nurilee.com/2020/04/03/lightgbm-definition-parameter-tuning/ (LGBM 사진)
[11] https://www.analyticsvidhya.com/blog/2022/01/histogram-boosting-gradient-classifier/ (histogram-based GBT 사진)