J. Integrative Natural Sci.
Vol.17, No.1 (2024) pp.1-12
https://doi.org/10.13160/ricns.2024.17.1.1

랜덤포레스트를 이용한 기상 환경에 따른 이상기온 분류

김윤수 · 송광윤 · 장인홍†

조선대학교 컴퓨터통계학과

Classification Abnormal temperatures based on Meteorological Environment using Random forests

Youn Su Kim, Kwang Yoon Song and In Hong Chang†

Department of Computer Science and Statistic, Chosun University, Gwangju, Korea

Abstract

Many abnormal climate events are occurring around the world. The cause of abnormal climate is related to temperature. Factors that affect temperature include excessive emissions of carbon and greenhouse gases from a global perspective, and air circulation from a local perspective. Due to the air circulation, many abnormal climate phenomena such as abnormally high temperature and abnormally low temperature are occurring in certain areas, which can cause very serious human damage. Therefore, the problem of abnormal temperature should not be approached only as a case of climate change, but should be studied as a new category of climate crisis. In this study, we proposed a model for the classification of abnormal temperature using random forests based on various meteorological data such as longitudinal observations, yellow dust, ultraviolet radiation from 2018 to 2022 for each region in Korea. Here, the meteorological data had an imbalance problem, so the imbalance problem was solved by oversampling. As a result, we found that the variables affecting abnormal temperature are different in different regions. In particular, the central and southern regions are influenced by high pressure (Mainland China, Siberian high pressure, and North Pacific high pressure) due to their regional characteristics, so pressure-related variables had a significant impact on the classification of abnormal temperature. This suggests that a regional approach can be taken to predict abnormal temperatures from the surrounding meteorological environment. In addition, in the event of an abnormal temperature, it seems that it is possible to take preventive measures in advance according to regional characteristics.

Keywords: Random forest, Imbalance data, Over-sampling, Climate crisis, Abnormal temperature

(**Recceived** February 8, 2024; **Revised** February 20, 2024; **Accepted** February 28, 2024)

† Corresponding author: ihchang@chosun.ac.kr

1. 서론

전 세계 각지에서 수많은 이상기후 사례들이 나타나고 있다. 그린란드 빙상 정상에 눈이 아닌 비가 내린 사례가 처음으로 기록됐고, 미국의 특정 주에서는 50도 이상의 기온이 관측됐으며, 중국의 한 지역에서는 수개월 동안의 강우량이 몇 시간 만에 내리는 폭우를 기록하기도 했다. 또한, 미국 캘리포니아 주, 호주의 대규모 산불 역시도 기후위기 사례 중 하나로 판단한다.

이러한 기후위기는 극심한 기후변화로부터 시작되었다. 기후변화는 자연적 원인과 인위적 원인으로 크게 두 가지로 나뉜다. 자연적 원인은 인간이 다룰 수 없는 영역으로 태양의 흑점 활동, 대규모 화산활동, 태양 복사 에너지의 변화로 인한 기온 변화 등이 있다. 인위적 원인은 인간이 문명을 발전시키면서 발생하는 탄소나 온실가스 농도 증가로 지구온난화가 가속화되었고, 해수면이 상승하여 단순 해안 저지대 침수나 홍수뿐만 아닌 강력한 태풍과 해일 등 큰 피해를 불러일으킬 자연재해를 동반한다[1].

기후변화에 가장 민감한 요소 중 기온을 상승시키는 요인은 여러 가지가 존재한다. 전 지구적인 관점으로는 온실가스와 탄소의 과다 배출로 인한 지구온난화를 가장 큰 요인으로 뽑는다. IPCC 보고서에 따르면 100년 전의 지구의 평균기온보다 평균기온이 1.09℃ 상승하였다[2]. 국부적 관점으로는 주변의 기압차에 의해 찬 공기와 따뜻한 공기의 순환으로 공기가 이동하며 각 위치의 기온이 변화하며, 극지방의 기류와 적도지방의 기류가 이동하며 기온이 변화한다. 이 중 기류에 의해 발생하는 기온 변화는 해당 지역에서 이상기후로 인한 극심한 피해를 일으킨다. 시베리아 지역에서 평균기온이 0℃인 지역에 38℃를 기록하거나 태국이나 인도, 이집트에서는 기록적 폭설이 내리기도 했다. 베트남에서는 이상저온 현상으로 난방시스템을 갖추고 있지 않아 많은 사상자를 낳은 것으로 보도되었다. 이러한 이상기후는 평균기온이 1.09℃ 증가했다는 수치로만 현상을 이해하기에는 부족한 부분이 많으며, 지구온난화로 평균기온이 상승한 것과는 별개의 극심한 피해를 끼치기 때문에 다른 기후위기의 관점으로 바라봐야한다.

지구온난화에 대한 문제는 대한민국도 마찬가지이다. 대한민국은 100년 전에 비해 약 1.6℃ 상승하여 전지구적 평균기온보다 더 높은 추이를 보였다. 이를 통해 과거에 위도가 낮은 지역에서 재배되던 농작물들의 재배지역이 북상하였고, 봄철 식물의 개화시기가 당겨지기도 하며, 해양 산성화 및 해수 온도 증가로 인해 해양생물이 북상하였다. 또한, 봄, 가을의 강수량 감소 및 겨울 적설량 감소로 인해 건조한 환경이 지속되면서 산불의 빈도가 많아지고, 여름철의 강수형태도 장마 이외에 태풍이나 집중호우의 빈도가 많아져, 과거보다 홍수나 산사태, 침수의 피해가 많아졌다. 또, 이상고온(폭염) 현상의 빈도 증가, 가뭄일수 증가로 농작물의 피해도 많아지고 있다[3]-[4].

기후위기의 심각성만큼 이상기후에 대한 연구가 여러 방면으로 이뤄지고 있다. 허인혜, 이승호[5]는 대한민국의 여름철과 겨울철의 이상기온 출현 빈도의 변화와 영향을 미치는 요인의 관계에 대한 연구를 진행하였다. 겨울 몬순 지수, 시베리아 고기압 강조 지수 및 북극 진동 지수, 오호츠크해 고기압 강도 지수, 북태평양 지수 등의 대한민국 주변의 기압 및 여러 지수들과 이상기온의 관계를 분석하여 변화 경향의 유사함을 입증하였다. 박미나, 최영은[6]은 14개의 기상관측지점, 고해상도 격자형 관측자료 기반 행정구역, RCP 시나리오 기반 미래 전망 기온 자료를 활용하여 최고기온 극값의 규모를 계산하여 변화 경향과 미래를 예측하였다. 해가 지날수록 최고기온이 꾸준히 상승하는 경향을 보였다. 이에 따라, 심창섭[7] 등은 여러 기후 시나리오 앙상블을 기반으로 지역별 기후변화 양상과 전망에 대해 논의하였으며, 열대야 일수는 증가하고, 남부지방의 고온다습한 기후는 심화될 것으로 예측하였고, 폭염빈도가 매우 증가할 것으로 내다봤다. 지역 차원의 기후변화 적응을 위해서는 지역 차원의 기후변화 특징을 파악하고 고해상도의 미래 기후전망이 필요할 것으로 보았다. 심교문 등[8]은 1973년부터 2010년까지 평균편차법을 통해 이상고온과 이상저온의 출현횟수를 조사하였고, 지역별로 분석하였다. 박태원 등[9]은 2000년부터 2019년까지의 일최고기온과 일최저기온 자료를 통해 한반도 폭염과 열대야 발생일수에 대한

해안과 내륙, 도시와 시골, 지리적 위치의 영향에 대해 상관분석을 통해 다방면의 관계를 분석하였다. 또한, 최근 머신러닝 기법을 활용하여 기상데이터를 활용한 연구가 진행되고 있다. 김영인 등[10]은 기계학습 방법 중 LSTM을 이용한 기온 및 폭염 발생 예측 방법론을 제시하여 33℃ 이상의 경우, 기존의 수치예보모형보다 제안한 모형의 정확도가 우수함을 입증하였고, 예측하기 위해 소요되는 시간 역시 매우 단축하였다.

본 연구에서는 2018년부터 2022년까지 이상기후 현상 중 이상기온에 초점을 두어 대한민국의 종관적 기상 특성 정보와 미세먼지 농도, 대한민국 상공의 1000hPa, 500hPa의 기상 정보, 탄소 농도에 대한 정보를 포함한 기상데이터를 활용하여 랜덤포레스트를 이용한 이상기온 분류 모형을 제안하고, 이에 따라 해마다 이상기온에 영향을 미치는 변수가 무엇인지 비교하고자 한다. 또한, 각 지역별 이상기온 변화에 영향을 미치는 주요변수를 비교하여, 지역별 특성을 고려한 영향 요인을 비교하고자 한다. 2절에서는 연구에 활용되는 데이터에 대한 소개, 3절에서는 본 연구에서 활용하는 연구 방법에 대한 소개, 4절에서는 수치적 예제를 통해 변수 중요도를 판단하고자 하며, 5장은 결론 및 제언으로 마무리한다.

2. 연구 방법

2.1. 랜덤포레스트

본 연구에서는 기계학습 방법 중 의사결정 나무 알고리즘과 배깅을 접목시킨 기계학습 방법인 랜덤포레스트(Random forest)를 활용한다[11]-[12]. 랜덤포레스트를 구성하는 의사결정나무는 지도학습 알고리즘에서 가장 유용하게 활용되는 분류와 예측 분석 방법이다. 의사결정하기 위한 추론 규칙이 나무와 같은 모양을 하고 있어 시각적으로 명료하게 표현된다. 의사결정 나무를 여러 개로 구성한 알고리즘이 랜덤포레스트이다. 이는 여러 의사결정나무의 결과를 종합하여 일반화 문제를 해결할 수 있다. 또한, 붓스트랩(Bootstrap) 방식을 활용하여 여러 데이터 세트를 구성하고 각각 학습시킨 후 결과를 결합하는 방법인 배깅을 통해 과적합이 발생하지 않도록 식 (1)을 통해 각 의사결정나무의 분산이 커지는 문제를 해결한다[13].

$$\hat{f}_{avg}(x)=\frac{1}{N}\sum_{n=1}^{N}\hat{f}^{*n}(x) \tag{1}$$

또한, 데이터를 통해 학습한 구조의 랜덤포레스트는 모형을 구성하는 여러 변수 중 종속변수에 영향을 가장 많이 미치는 변수가 무엇인지 판단하는 변수 중요도를 계산한다. 랜덤포레스트에서 변수의 중요도를 판단하는 지니 불순도는 모형을 구성하는 특정 변수가 모형에 적용될 때, 분류시키는데 있어 얼마나 불순도를 감소시키는지를 나타낸다. 지니 불순도는 랜덤포레스트에서 각 노드마다의 불순도를 의미하며 둘의 차이가 클수록 변수 중요도가 높은 것으로 판단하며, 식 (2)를 따른다[14]-[15].

$$G(X)=-\sum_{n=1}^{N}p(x_n)(1-p(x_n)) \tag{2}$$

2.2. 오버샘플링

불균형 데이터는 얻어진 클래스의 비율이 하나에 매우 치중된 데이터이다. 예를 들어, 고장인지 아닌지 분류할 때, 고장의 비율이 전체 비중 중 약 10% 미만이라면 이를 분류 분석을 진행할 때, 정분류율은 높을 수 있지만, 고장에 대한 클래스를 올바르게 분류하지 못하는 문제가 발생한다. 이와 같은 클래스 불균형 문제를 해결하기 위한 방법에는 소수 클래스를 다수 클래스 크기만큼 늘리는 방법인 오버샘플링(Oversampling) 방법과 다수 클래스를 소수 클래스 크기에 맞추는 방법인 언더샘플링(Undersampling) 방법이 존재한다. 이 중 본 연구에서는 오버샘플링을 활용하여 소수 클래스를 복제한 후, 소수 클래스 크기를 랜덤포레스트 모형에 학습시키고자 한다. 소수 클래스를 그대로 복제시켰기 때문에 과적합 문제가 발생할 수 있으니 다수 클래스 크기만큼 복제하지 않고, 전체 데이터 수 중 약 30%에 해당하는 만큼 오버샘플링하여 분석에 활용하고자 한다[16]. 학습된 모형을 정분류율, 민감도, 특이도를 평가지표로서 활용한다. 여기서, 민감도는

“소수 클래스로 정확하게 예측한 빈도수 / 실제 소수 클래스 빈도수”이고, 특이도는 “다수 클래스로 정확하게 예측한 빈도수 / 실제 다수 클래스 빈도수”를 의미한다[17]. 본 연구에서는 정분류율(Accuracy, ACC), 민감도(Sensitivity, SEN), 특이도(Specificity, SPE) 세 가지의 평가지표를 활용하며, 전체 데이터 세트 중 70%는 훈련용 데이터, 30%는 테스트용 데이터로 나누어 분석한다.

3. 데이터 소개

본 연구에서는 기상청에서 제공하는 종관기상관측 자료, 황사 관측 자료, 레존윈데 자료, 자외선 자료, 온실가스와 반응가스 자료를 통해 이상고온과 이상저온을 포함하는 이상기온에 대한 분류를 위해 활용한다.

3.1. 종관기상관측

종관규모의 날씨를 파악하기 위해 정해진 시간에 모든 관측소에서 같은 시각에 측정을 실시하는 지상관측을 의미한다. 여기서 종관규모는 고기압이나 저기압의 크기 및 수명을 의미하고, 기온, 강수, 강설, 바람, 습도, 기압, 일조, 일사, 구름 등과 같은 매일의 날씨 현상을 102개의 관측소에서 측정하고, 기록하여 제시한다. 본 연구에서는 관측지역의 기상 상황을 활용하고자 하며, 활용하는 변수는 강수지속시간, 일강수량, 최대풍속, 평균풍속, 평균이슬점온도, 평균상대습도, 평균현지기압, 평균해면기압, 합계일사량이다. 평균기온 및 최고, 최저기온은 이상기온과 매우 직접적 관련이 있는 자료이기 때문에 이는 제외한다.

3.2. 황사 관측

부유분진측정기(PM10)는 대기 중에 부유하는 에어로졸 중 직경이 10㎛ 이하인 입자를 먼지 필터에 침적시킨 후, 동위원소 C-14에서 방출되는 베타선을 필터 여지에 투영시켜 감쇄된 베타선을 검출기로 측정하며, 총 31개 지점에서 관측하여 각 지역의 미세먼지 농도를 관측한다.

3.3. 레존윈데

라디오존데를 기구에 달아 비양시켜 상공 30km까지 대기상태를 일정 간격으로 총 10개의 지점에서 관측하며, 오전, 오후 12시간 간격으로 측정한다. 본 연구에서 활용하고자 하는 변수는 등기압 기준으로 0시 12시 시점의 500㎪, 1000㎪이 기록된 고도, 풍속이며, 낮은 기압의 정보는 공기의 순환에 영향을 미치며, 궁극적으로 기온에 영향을 미치는 변수이기 때문에 이상기온 분류에 활용하고자 한다.

3.4. 자외선

자외선은 일반적으로 자외선A, 자외선B, 자외선C로 나뉘며, 기상청에서 관측하는 자외선은 자외선A(320~400nm)에 대한 일누적자외선A와 자외선B 영역 중 인체에 홍반을 발생시키는 홍반자외선B(280~320nm)에 대한 일최대자외선B를 관측한다. 자외선을 측정하는 관측지점은 안면도, 울릉도·독도이며, 본 연구에서는 해당 지역에서 관측한 자료를 중심으로 일누적자외선A와 일최대자외선B를 이상기온 분류에 활용하고자 한다.

3.5. 온실가스 및 반응가스

온실가스는 태양열을 지구로 투과시키고, 지표에서 방출되는 열을 흡수하거나 재방출하여 지구의 온도를 유지하도록 도와주는 역할을 하는 물질을 의미하고, 반응가스는 체류시간이 짧지만 다른 물질들과의 반응성이 높아 반응가스라고 불린다. 반응가스 중 일산화탄소와 질소산화물의 영향으로 생성되는 지표오존은 다른 온실가스와 함께 강력한 온실가스를 보인다. 이 중 화석연료나 산화과정 등으로 발생하는 일산화탄소는 수산화이온과의 반응으로 이산화탄소 등 온실가스 농도를 변화시켜 지구온난화에 영향을 미친다고 알려져 있다[18]. 따라서 본 연구에서는 일산화탄소를 이상기온 분류에 활용하고자 하며, 가장 많이 알려진 이산화탄소나 메탄 등은 결측치로 포함된 자료가 많기 때문에 온실가스 대신 일산화탄소로 대신하여 분석에 활용한다.

3.6. 이상기상

이상기상은 기온, 강수량 등의 기상요소가 평년에 비해 현저히 높거나 낮은 수치를 나타내는 극한 현상을 의미하며, 이상기상을 정의하는 방법은 국립원예특작과학원에서 측정하는 방식인 정규분포를 활용하는 확률론적인 방법과 기상청에서 측정하는 방식인 백분위수를 활용하는 퍼센타일 방법이 있다. 본 연구에서는 퍼센타일 방법을 활용하여 이상기상을 판단하며, 기온, 강수 등 다양한 기상요소 중 기온만을 가지고 분석에 활용한다. 이상저온은 최저·최고기온 10백분위수 미만, 이상고온은 90백분위수 초과로 정의하며, 본 연구에서는 이상저온과 이상고온을 통합하여 이상기온으로 활용하고자 한다[5].

4. 수치적 예제

4.1. 기술통계

표 1은 지역별 종관기상관측 및 황사 관측으로 측정된 데이터에 대한 평균 및 표준편차이다. 강수지속시간이 가장 긴 지역은 경북으로 나타났고, 일강수량, 최대풍속, 평균풍속, 평균이슬점온도, 평균상대습도, 평균현지기압은 제주에서 가장 높게 나타났다. 최고해면기압은 강원에서 가장 높았고, 최저해면기압과 평균해면기압은 경기와 전북, 충남에서 가장 높게 나타났고, 합계일사량은 충북에서 가장 높게 나타났으며 미세먼지 농도가 가장 높은 지역은 경기로 나타났다.

표 2는 반응가스와 자외선 양에 대한 평균과 표준편차를 제시하였다. 일산화탄소는 안면도와 백령도에서 측정한 결과가 가장 높은 것으로 나타났고, 일누적자외선A는 안면도에서 높게 나타났고, 일최대자외선B는 울릉도와 안면도가 동일하게 나타났다. 일산화탄소를 측정한 지역은 경북의 울릉도, 충남의 안면도, 제주의 고산이고, 자외선을 측정한 지역은 경북의 울릉도, 충남의 안면도이다. 해당 지역을 제외한 다른 지역은 인접지역의 일산화탄소와 일누적 및 일최대자외선B을 분석에 활용한다.

Table. 1. 종관기상관측과 황사 관측에 대한 기술통계량 결과

지역	강수지속시간(hr)		일강수량(㎜)		최대풍속(㎧)		평균풍속(㎧)		평균이슬점온도(℃)	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
강원	2.89	4.78	4.41	10.97	4.25	1.23	1.62	0.63	4.87	12.02
경기	3.09	5.56	4.12	12.09	4.62	1.32	2.02	0.75	6.34	11.73
경남	2.10	4.20	4.68	14.29	4.23	1.17	1.67	0.63	7.61	11.49
경북	3.48	4.95	4.61	10.38	4.77	1.36	2.01	0.77	6.26	11.62
전남	2.53	4.13	4.28	12.88	5.58	1.59	2.47	1.03	9.00	10.36
전북	2.39	4.50	4.19	12.57	4.58	1.24	1.70	0.70	7.88	10.60
제주	2.75	4.64	4.92	15.36	6.52	2.11	3.54	1.37	11.83	8.98
충남	2.61	4.79	3.95	12.30	3.94	1.11	1.45	0.63	7.26	11.08
충북	2.39	4.68	3.94	11.56	4.21	1.18	1.58	0.71	5.72	11.51
전체	2.69	4.72	4.34	12.58	4.75	1.60	2.01	1.04	7.42	11.25
지역	평균상대습도(%)		평균현지기압(hPa)		평균해면기압(hPa)		합계일사량(MJ/㎡)		미세먼지농도(㎍/㎥)	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
강원	66.94	15.09	990.94	7.23	1016.34	8.17	14.32	6.84	27.66	18.06
경기	69.61	13.87	1009.30	8.15	1016.73	8.48	14.40	6.93	35.30	22.50
경남	67.39	15.94	1008.59	7.50	1016.24	7.76	14.46	6.78	26.26	16.43
경북	67.00	15.70	1002.05	7.41	1016.27	7.92	14.30	6.76	25.92	17.35
전남	72.84	12.72	1009.44	7.90	1016.41	7.98	14.46	6.78	28.10	21.77
전북	72.91	11.50	1002.56	7.77	1016.72	8.31	13.74	6.75	33.44	25.20
제주	74.30	12.64	1011.44	7.60	1016.27	7.74	13.67	7.48	29.23	27.59
충남	71.43	12.47	1009.48	8.09	1016.72	8.36	14.76	7.18	32.02	22.87
충북	67.63	13.47	996.53	7.49	1016.59	8.37	14.88	7.19	30.14	19.86
전체	70.01	14.05	1004.49	10.14	1016.48	8.13	14.33	6.98	29.80	21.67

Table. 2. 반응가스와 자외선에 대한 기술통계량 결과

지역	일산화탄소(ppb)		일누적자외선 A(MJ/m²)		일최대자외선 B(W/m²)	
	mean	sd	mean	sd	mean	sd
울릉도	174.24	51.41	0.70	0.41	0.13	0.07
안면도	254.89	108.36	0.77	0.39	0.13	0.07
고산	187.86	72.41	-	-	-	-
전체	205.21	87.87	0.75	0.40	0.13	0.07

레존윈데를 관측하는 장소는 강원도의 북강릉, 경기도의 백령도, 오산, 전남의 광주, 흑산도, 경남의 창원, 경북의 포항으로 표 3은 각 지역을 대표하는 자료를 기반으로 측정한 시간 및 기압에 따른 고도, 풍속에 대한 평균과 표준편차를 보여준다. 0시와 12시의 500hPa 고도는 전남지역에서 가장 높았으며, 0시와 12시의 1000hPa 고도는 경기지역에서 가장 높게 나타났다. 0시의 500hPa 풍속은 경북에서 가장 높게 나타났고, 12시 1000hPa 풍속은 경기도에서 가장 높게 나타났으며 0시의 1000hPa 풍속과 12시 1000hPa 풍속은 경북과 경기에서 각각 가장 높게 나타났다. 전북, 제주, 충남, 충북에는 관측 장소가 없기 때문에 인접지역의 값을 분석에 활용한다.

표 4는 2018년부터 2022년까지 대한민국 전 지역의 이상기온 발생일의 정보를 포함한다. 5년간의 이상기온은 2018년 가장 많이 발생하였고, 2019년에 가장 적게 발생하였으며 다시 상승하는 추세를 보이고 있다. 대체적으로 이상기온은 충남, 충북 지역에서 가장 많이 발생하였고, 강원과 경북지역은 이상기온이 적게 발생한 지역으로 나타났다. 또한, 2019년과 2022년은 제주지역도 이상기온 현상이 많이 발생하였다.

Table. 3. 레존윈데 관측에 대한 기술통계량 결과

지역	0시 500hPa 고도		0시 500hPa 풍속		0시 1000hPa 고도		0시 1000hPa 풍속	
	mean	sd	mean	sd	mean	sd	mean	sd
강원	5650.80	160.28	40.58	21.58	148.43	74.22	5.37	3.90
경기	5659.89	156.73	39.53	20.47	153.27	73.65	8.66	4.70
경남	5701.44	136.48	41.08	21.27	148.94	74.03	5.95	4.34
경북	5684.45	144.96	41.53	21.50	146.06	69.31	10.76	5.75
전남	5707.30	131.44	39.66	20.73	147.11	70.15	8.41	5.26
전체	5682.02	147.40	40.08	20.90	149.41	72.17	8.20	5.14
지역	12시 500hPa 고도		12시 500hPa 풍속		12시 1000hPa 고도		12시 1000hPa 풍속	
	mean	sd	mean	sd	mean	sd	mean	sd
강원	5654.78	161.80	40.22	22.09	143.16	78.56	5.69	4.05
경기	5664.00	159.43	39.78	21.40	147.24	80.51	11.77	7.03
경남	5705.64	135.77	40.84	22.02	141.83	72.47	5.69	3.94
경북	5687.39	145.79	41.05	22.33	139.52	69.31	10.09	5.61
전남	5710.03	131.68	39.47	21.14	143.90	75.71	9.50	5.40
전체	5685.54	148.55	39.98	21.56	144.20	76.64	9.50	6.15

Table. 4. 이상기온에 대한 빈도분석 결과

지역	2018년		2019년		2020년		2021년		2022년	
	보통	이상기온	보통	이상기온	보통	이상기온	보통	이상기온	보통	이상기온
강원	337	28	357	8	341	24	349	16	348	17
경기	305	60	332	33	316	49	298	67	318	47
경남	300	65	336	29	326	39	324	41	312	53
경북	314	51	342	23	331	34	337	28	335	30
전남	306	59	340	25	325	40	315	50	317	48
전북	307	58	334	31	317	48	313	52	318	47
제주	308	57	323	42	313	52	303	62	303	62
충남	285	80	329	36	307	58	289	76	300	65
충북	287	78	321	44	311	54	299	66	304	61
전체	305.4	59.6	334.9	30.1	320.8	44.2	314.1	50.9	317.2	47.8

4.2. 분류율

표 5와 6은 지역별 2018년부터 2022년까지의 랜덤포레스트를 활용한 기상데이터를 활용한 이상기온 분류 모형의 훈련용(Trian) 데이터 70%와 검증용(Test) 데이터 30%에 대한 분석 결과이다. 이 역시도 훈련용 데이터와 검증용 데이터의 분류율이 92~98%로 나타나 상당히 좋은 결과를 보였으며, 과적합 발생한 결과는 나타나지 않았다. 오버샘플링을 통해 불균형 데이터의 문제인 소수 클래스 분류에 대한 특이도도 좋은 결과를 보였다.

4.3. 변수 중요도

표 7은 2018년부터 2022년까지의 경기·강원·경남 지역의 이상기온 분류에 영향을 많이 미치는 상위

Table. 5. 훈련용 데이터 세트의 정분류율, 민감도, 특이도

년도	Train	경기	강원	경남	경북	충남	충북	전남	전북	제주
2018	Accuracy	95.6%	98.2%	94.9%	97.4%	94.3%	93.4%	96.1%	96.4%	94.3%
	Sensitivity	94.8%	98.0%	93.9%	97.1%	91.2%	89.8%	95.5%	95.3%	92.8%
	Specificity	97.0%	99.3%	96.7%	98.3%	98.0%	97.9%	97.1%	98.1%	96.7%
2019	Accuracy	95.7%	98.3%	95.0%	96.9%	94.5%	92.9%	95.8%	96.9%	94.4%
	Sensitivity	94.6%	98.2%	94.1%	96.5%	91.4%	88.9%	95.3%	96.1%	92.9%
	Specificity	97.4%	99.3%	96.7%	97.8%	98.1%	97.8%	96.7%	98.2%	96.7%
2020	Accuracy	95.6%	98.2%	95.0%	96.8%	94.3%	92.7%	95.8%	96.3%	94.4%
	Sensitivity	94.8%	98.0%	94.1%	96.2%	91.2%	88.4%	95.8%	95.0%	93.0%
	Specificity	97.0%	99.3%	96.7%	98.3%	98.0%	97.9%	96.9%	98.4%	96.5%
2021	Accuracy	95.9%	98.3%	95.0%	97.1%	94.5%	93.0%	96.6%	96.4%	94.3%
	Sensitivity	95.0%	98.2%	94.2%	96.6%	91.6%	89.3%	96.4%	95.3%	92.8%
	Specificity	97.2%	99.3%	96.6%	98.3%	98.0%	97.6%	97.1%	98.2%	96.7%
2022	Accuracy	95.9%	98.3%	95.0%	96.9%	94.3%	92.9%	96.4%	96.3%	94.4%
	Sensitivity	94.9%	98.2%	94.2%	96.4%	91.3%	89.0%	96.2%	95.2%	93.0%
	Specificity	97.4%	99.3%	96.6%	98.0%	98.0%	97.8%	96.7%	98.1%	96.5%

Table. 6. 검증용 데이터 세트의 정분류율, 민감도, 특이도

년도	Test	경기	강원	경남	경북	충남	충북	전남	전북	제주
2018	Accuracy	94.0%	97.7%	96.0%	96.8%	93.6%	94.8%	94.9%	94.6%	97.4%
	Sensitivity	92.2%	98.3%	93.9%	96.3%	89.1%	91.8%	95.1%	94.2%	95.4%
	Specificity	96.9%	95.0%	100.0%	98.0%	100.0%	98.5%	94.5%	95.5%	100.0%
2019	Accuracy	93.7%	97.9%	96.3%	96.8%	93.5%	94.6%	94.9%	94.9%	97.2%
	Sensitivity	91.7%	98.5%	94.4%	96.3%	88.9%	91.5%	95.1%	94.6%	95.1%
	Specificity	96.9%	95.0%	100.0%	98.0%	100.0%	98.5%	94.5%	95.5%	100.0%
2020	Accuracy	93.7%	97.7%	96.3%	97.1%	93.6%	93.9%	95.5%	94.5%	97.2%
	Sensitivity	91.7%	98.3%	94.4%	96.7%	89.1%	90.3%	95.3%	94.0%	95.1%
	Specificity	96.9%	95.0%	100.0%	98.0%	100.0%	98.5%	95.9%	95.5%	100.0%
2021	Accuracy	94.0%	97.9%	96.3%	96.9%	93.2%	94.5%	95.1%	95.0%	97.2%
	Sensitivity	92.2%	98.5%	94.4%	96.5%	88.4%	91.2%	95.5%	94.0%	95.1%
	Specificity	96.9%	95.0%	100.0%	98.0%	100.0%	98.5%	94.5%	97.0%	100.0%
2022	Accuracy	93.5%	97.9%	96.2%	96.9%	93.6%	94.1%	95.1%	95.3%	97.6%
	Sensitivity	92.4%	98.5%	94.2%	96.5%	89.1%	90.6%	95.5%	94.4%	95.7%
	Specificity	95.3%	95.0%	100.0%	98.0%	100.0%	98.5%	94.5%	97.0%	100.0%

7개 변수에 대한 결과이다. 경기 지역은 0시 1000 hPa 기온, 미세먼지농도, 평균이슬점온도, 0시 500hPa 풍속이 이상기온에 영향을 많이 미치는 변수로 나타났고, 2018년에는 강수지속시간이 0시 1000hPa 풍속보다 더 높은 영향을 미쳤지만, 2019년부터 2021년까지는 0시 1000hPa 풍속이 강수지속시간보다 더 높은 영향을 미쳤으며, 2022년에는 다시 2018년과 동일하게 영향을 미쳤다. 7번째로 높게 영향을 미친 변수는 12시 1000hPa 기온으로 나타났다. 지역 특성상 중국 대륙, 시베리아 부근에서 발생한 고기압으로부터 영향을 많이 받아 500hPa의 풍속에 대한 영향이 다른 지역에 비해 높은 것으로 보인다. 강원 지역은 2018년부터 2022년까지 1순위부터 5순위까지 동일한 변수가 영향을 미치는 것으로 나타났고, 2018년은 합계일사량이 일누적자외선A보다 높은 영향을 미치는 것으로 나타났고, 2019년부터 2022년까지는 일누적자외선A가 합계일사량보다 높게 영향을 미치는 것으로 나타났다. 산간지역이 많은 강원이기 때문에 평균기온이 다른 지역에 비해 낮아 겨울철 이상저온에 영향을 많이 미치는 것으로 보인다. 경남 지역은 0시 500hPa 풍속, 12시 1000hPa 기온, 평균이슬점온도, 0시 1000hPa 기온, 일산화탄소

Table. 7. 경기·강원·경남 지역의 상위 7개 주요 변수

경기	2018	2019	2020	2021	2022
1	미세먼지농도	미세먼지농도	미세먼지농도	0시 500hPa 고도	미세먼지농도
2	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도	미세먼지농도	0시 500hPa 고도
3	평균이슬점온도	강수지속시간	강수지속시간	강수지속시간	강수지속시간
4	강수지속시간	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
5	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속
6	0시 1000hPa 풍속	0시 1000hPa 풍속	0시 1000hPa 풍속	0시 1000hPa 풍속	0시 1000hPa 풍속
7	12시 1000hPa 고도	12시 1000hPa 고도	12시 1000hPa 고도	12시 1000hPa 고도	12시 1000hPa 고도
강원	2018	2019	2020	2021	2022
1	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
2	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도
3	일산화탄소	일산화탄소	일산화탄소	일산화탄소	일산화탄소
4	합계일사량	합계일사량	합계일사량	합계일사량	합계일사량
5	0시 500hPa 풍속	12시 1000hPa 고도	0시 500hPa 풍속	0시 500hPa 풍속	12시 1000hPa 고도
6	12시 1000hPa 고도	0시 500hPa 풍속	12시 1000hPa 고도	12시 1000hPa 고도	0시 500hPa 풍속
7	0시 1000hPa 고도	0시 1000hPa 풍속	0시 1000hPa 고도	0시 500hPa 고도	0시 1000hPa 고도
경남	2018	2019	2020	2021	2022
1	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
2	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속
3	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도
4	일산화탄소	일산화탄소	일산화탄소	일산화탄소	일산화탄소
5	최대풍속	최대풍속	최대풍속	최대풍속	최대풍속
6	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A
7	합계일사량	합계일사량	평균풍속	평균풍속	합계일사량

순으로 영향을 많이 미치는 것으로 나타났고, 2018년도는 최대풍속이, 2019년부터 2022년까지는 미세먼지 농도가 더 높게 나타났다. 북태평양 고기압으로부터 여름철 고온다습한 공기의 유입으로 이상기온에 영향을 많이 미치는 것으로 보인다.

표 8은 2018년부터 2022년까지의 경북·충남·충북 지역의 이상기온 분류에 영향을 많이 미치는 상위 7개 변수에 대한 결과이다. 경북 지역은 12시 1000 hPa 기온, 미세먼지농도가 이상기온 분류에 영향을 많이 미친 변수로 나타났고, 2019년은 3순위가 평균이슬점온도, 4순위가 일산화탄소로 나타났고, 이를 제외한 2018년, 2020년부터 2022년까지는 모두 동일한 순서대로 이상기온 분류에 영향을 미쳤다. 다른 지역에 비해 일산화탄소의 농도와 미세먼지 농도가 영향을 많이 끼치는 것을 확인할 수 있다. 충남 지역은 2018년부터 2022년까지 동일하게 0시 1000hPa 기온, 0시 500hPa 고도, 평균이슬점온도, 12시 1000hPa 기온, 미세먼지농도, 일산화탄소, 일누적자외선A 순으로 이상기온 분류에 영향을 미치는 것으로 나타났다. 충북 지역도 마찬가지로 2018년부터 2022년까지 동일하게 평균이슬점온도, 0시 1000 hPa 기온, 0시 500hPa 고도, 일산화탄소, 12시 1000

Table. 8. 경북·충남·충북 지역의 상위 7개 주요 변수

경북	2018	2019	2020	2021	2022
1	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
2	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도
3	일산화탄소	일산화탄소	일산화탄소	일산화탄소	일산화탄소
4	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도
5	12시 1000hPa 고도	12시 1000hPa 고도	12시 1000hPa 고도	12시 1000hPa 고도	12시 1000hPa 고도
6	평균상대습도	평균상대습도	평균상대습도	평균상대습도	평균상대습도
7	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속
충남	2018	2019	2020	2021	2022
1	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
2	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도
3	일산화탄소	일산화탄소	일산화탄소	일산화탄소	일산화탄소
4	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A
5	평균상대습도	0시 1000hPa 풍속	평균상대습도	0시 1000hPa 풍속	평균상대습도
6	0시 1000hPa 풍속	평균상대습도	미세먼지농도	미세먼지농도	0시 1000hPa 풍속
7	미세먼지농도	미세먼지농도	0시 1000hPa 풍속	평균상대습도	미세먼지농도
충북	2018	2019	2020	2021	2022
1	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
2	일산화탄소	일산화탄소	일산화탄소	일산화탄소	일산화탄소
3	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 풍속	0시 500hPa 고도	0시 500hPa 고도
4	0시 1000hPa 풍속	0시 1000hPa 풍속	0시 500hPa 고도	0시 1000hPa 풍속	0시 500hPa 풍속
5	0시 500hPa 풍속	0시 500hPa 풍속	0시 1000hPa 풍속	0시 500hPa 풍속	0시 1000hPa 풍속
6	일최대자외선B	일최대자외선B	일최대자외선B	일최대자외선B	일최대자외선B
7	미세먼지농도	미세먼지농도	일누적자외선A	미세먼지농도	일누적자외선A

hPa 기온, 0시 500hPa 풍속, 0시 1000hPa 풍속 순으로 이상기온 분류에 영향을 미치는 것으로 나타났다. 충남·충북 지역은 수도권인 경기 지역과 인접 지역으로 대륙 고기압에 영향을 받아 기압 관련 변수에 영향을 많이 받는 것으로 보인다.

표 9는 2018년부터 2022년까지의 전남·전북·제주 지역의 이상기온 분류에 영향을 많이 미치는 상위 7개 변수에 대한 결과이다. 전남 지역은 2018년부터 2022년까지 일산화탄소, 평균이슬점온도, 12시 1000hPa 기온이 이상기온 분류에 높게 영향을 미치는 변수로 나타났고, 2018년도는 미세먼지가 4번째로 영향을 많이 미치는 변수로 나타났으나 2019년부터 2022년까지는 평균상대습도가 4번째로 영향을 많이 미치는 것으로 나타났다. 또, 5,6,7번째 변수는 평균상대습도, 미세먼지 농도, 12시 1000hPa 풍속, 0시 1000hPa 기온, 일강수량의 변수가 바뀌어가며 영향을 많이 미치는 것으로 나타났다. 전남 지역은 넓은 부지의 공장단지가 조성되어있어 해당 지역에서 발생하는 일산화탄소가 가장 많은 영향을 미치는 것으로 보인다. 전북 지역은 2018년부터 2022년까지 평균이슬점온도, 0시 1000hPa 기온이 이상기온 분류에 높게 영향을 미치는 변수로 나타났고, 평균풍속

Table. 9. 전남·전북·제주 지역의 상위 7개 주요 변수

전남	2018	2019	2020	2021	2022
1	평균상대습도	평균상대습도	평균상대습도	평균상대습도	평균상대습도
2	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
3	일산화탄소	일산화탄소	일산화탄소	일산화탄소	일산화탄소
4	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도
5	평균풍속	평균풍속	평균풍속	평균풍속	평균풍속
6	0시 500hPa 풍속	12시 1000hPa 풍속	12시 1000hPa 풍속	12시 1000hPa 풍속	12시 1000hPa 풍속
7	12시 1000hPa 풍속	0시 500hPa 풍속	0시 500hPa 풍속	일누적자외선A	0시 500hPa 풍속
전북	2018	2019	2020	2021	2022
1	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
2	평균풍속	평균풍속	평균풍속	평균풍속	평균풍속
3	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A
4	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도
5	일최대자외선B	일산화탄소	일최대자외선B	일산화탄소	일산화탄소
6	일산화탄소	일최대자외선B	일산화탄소	일최대자외선B	일최대자외선B
7	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도	0시 500hPa 고도
제주	2018	2019	2020	2021	2022
1	평균상대습도	평균상대습도	평균상대습도	평균상대습도	평균상대습도
2	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도	평균이슬점온도
3	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A	일누적자외선A
4	일산화탄소	일산화탄소	일산화탄소	일산화탄소	일산화탄소
5	일최대자외선B	일최대자외선B	일최대자외선B	일최대자외선B	일최대자외선B
6	12시 1000hPa 풍속	12시 1000hPa 풍속	12시 1000hPa 풍속	12시 1000hPa 풍속	12시 1000hPa 풍속
7	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도	미세먼지농도

과 일산화탄소, 미세먼지농도가 3, 4, 5순위에 위치해있다. 6번째, 7번째로 영향을 많이 미친 변수는 일누적자외선A과 일최대자외선B으로 나타났다. 인접지역이 전남지역이며, 해당 지역도 넓은 부지의 공장이 조성되어 있기 때문에 일산화탄소나 미세먼지 농도에 영향을 많이 미치는 것으로 보인다. 제주지역은 평균상대습도가 가장 영향을 많이 미치는 변수로 나타났고, 일누적자외선A과 평균이슬점온도, 일산화탄소가 2, 3, 4순위에 위치해있으며 5번째로 높은 변수는 최고해면기압으로 나타났다. 이는 다른 지역에 비해 사방이 바다로 둘러 쌓여있기 때문에 가장 큰 영향을 미치는 것으로 보인다.

5. 결론 및 제언

본 연구에서는 지역별로 관측된 기상데이터를 활용하여 이상기상 중 하나인 이상기온을 랜덤포레스트를 통해 분류하였으며, 이상기온을 분류함에 있어서 영향을 가장 많이 끼치는 변수를 측정하였다. 활용한 기상데이터는 불균형데이터의 특징을 지니고 있기 때문에 오버샘플링을 통해 불균형 문제를 해결하여 분석하였다. 그 결과, 지역별로 이상기온에 영향을 미치는 변수가 다르다는 결과를 얻었다. 중부지역과 남부지역은 지역적 특성상 고기압(중국대륙, 시베리아 고기압 및 북태평양 고기압)의 영향을 받아 기압 관련 변수가 이상기온 분류에 큰 영향을 미친 것으로 나타났고, 공장 부지가 있는 지역은 일산화탄소와 미세먼지의 농도가 큰 영향을 미치는 것으로 나타났으며, 중국과 인접해있는 지역은 미세먼지 농도의 영향이 높게 나타났다. 이를 통해 지역적 특색에 따라 접근하여 주위의 여러 기상 환경을 통해 이상기온을 예측할 수 있다. 또한, 이상기온이 발생했을 경우, 지역적 특색에 맞게 미리서 예방할 수 있는 대비책을 마련할 수 있을 것으로 보인다.

Acknowlegements

이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019S1A6A3A01059888).

참고문헌

[1] 임인재, 박성철, 이덕환., 논쟁적 과학이슈에 대한 신문보도 분석: 기후변화의 원인과 대응 관련 보도를 중심으로. 한국언론학보, Vol. 57, No. 6, pp. 469-501, 2013.

[2] 김준, et al., Korean Climate Change Assessment Report 2020. 2020.

[3] 국립기상과학원, 한반도 100년의 기후변화, 2018

[4] 한국. 2022년 이상기후 보고서. 서울: 기상청, 2023.

[5] 허인혜, 이승호., 한국의 이상기온 출현 빈도의 변화와 그 요인에 관한 연구. 대한지리학회지, Vol. 41, No. 1, pp. 94-105, 2006.

[6] 박미나, 최영은., 우리나라의 재현기간별 일 최고기온 극값 변화 및 미래 전망에 관한 연구. 국토지리학회지, Vol. 54, No. 1, pp. 37-48, 2020.

[7] 심창섭, 문준기, 한지현, 서지현, 송영일, 홍제우, 유명수., 우리나라 지역별 기후변화 전망과 적응정책 차원의 시사점., 2022.

[8] 심교문, 김용석, 정명표, 김석철, 민성현, 소규호., 한국의 농업기후지대별 이상기온 출현 특성 평가. 한국기후변화학회지, Vol. 4, No. 2, pp. 189-199, 2013.

[9] 박태원, 박현빈, 장준기, 박종석, 유민숙., 한반도에서의 이상고온 발생의 지역적 특징. 과학영재교육, Vol. 13, No. 1, pp. 1-12, 2021.

[10] 김영인, 김동현, 이승오., 기계학습을 활용한 하절기 기온 및 폭염발생여부 예측. 한국방재안전학회 논문집, Vol. 13, No. 2, pp. 27-38, 2020.

[11] Breiman, L., Random forests. Machine learning, Vol. 45, pp. 5-32, 2001.

[12] Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H., "The elements of statistical learning: Data mining, inference, and prediction, Springeer, New York.", 2009.

[13] Park, E. J., Park, J. H., & Kim, H. H., "Mapping species-specific optimal plantation sites using random forest in Gyeongsangnam-do province, South Korea.", J. Agric. Life Sci, Vol. 53, pp. 65-74, 2019.

[14] Qiu, X. and Choi, P., "A study on discrimination in mortgage lending in the United States: A revisit by random forest method.", Journal of the Korean Data & Information Science Society, Vol. 30, pp. 261-370, 2019.

[15] Heo, T. I., Kim, D. H., and Hwang, S. W., "Identification of Celtis species using random forest with infrared spectroscopy and analysis of spectral feature importance.", Journal of the Korean Data & Information Science Society, Vol. 32, pp. 1183-1194, 2021.

[16] Kim, H., & Lee, W., "On sampling algorithms for imbalanced binary data: performance comparison and some caveats.", The Korean Journal of Applied Statistics, Vol. 30, No. 5, pp. 681-690, 2017.

[17] Hong, C. S., and Jang, D. H., "Partial AUC using the sensitivity and specificity lines.", Korean Journal of Applied Statistics, Vol. 33, pp. 541-553, 2020.

[18] Stocker, T. (Ed.)., "Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change.", Cambridge university press., 2014.