


<div>  <div> <div>〈2023 날씨 빅데이터 콘테스트〉</div> <div>기상에 따른 계절별 지면온도 산출 기술 개발 방안</div> </div> </div>			
참 가 번 호	230043	팀 명	경희의 온도🌡

1. 분석배경 및 목표

지면온도는 농업, 에너지, 건축, 환경, 국방 등 수많은 분야에서 유용하게 활용되고 있으며, 기후변화, 농작물의 생육 조건 파악, 지열 발전의 잠재력 평가, 지구 온난화 현상 추적, 가뭄 감시 등 연구적 측면에서까지 중요한 기상인자로 사용되고 있는 실정이다. 이러한 실질적 수요에 입각하여 지면온도를 정확히 파악하고 데이터를 체계적으로 구축하는 것은 사회적, 경제적 측면에서 대단히 중요한 성과로 이어질 수 있다. 따라서 본 콘테스트에서는 시공간적으로 상세한 계절별 지면온도 산출 기술을 개발하여 양질의 지면온도 데이터를 확보하는 것이 주된 목적이다. 이에 따라 경희의 온도팀은 기상 빅데이터를 통해 계절별 지면온도 예측과 밀접한 관련이 있는 특징들을 탐색하고 상황에 맞는 파생변수를 만들어 모델의 정확도와 설명력을 높이는 것을 목표로 하고자 한다.

2. 데이터 정의 및 EDA

2.1. 데이터 정의

〈표 1〉 Train set 변수 개요

변수명	정의	변수명	정의	변수명	정의
STN	지점번호(1~10)	HM	1시간 평균 상대습도(%)	SI	1시간 누적 일사량(MJ)
YEAR	년도(A, B, C, D, E)	WS	1시간 평균 풍속(m/s)	SS	1시간 누적 일조량(초)
MMDDHH	월/일/시간	RN	1시간 누적 강수량(mm)	SN	00분에 측정된 적설 깊이(cm)
TA	1시간 평균 기온(℃)	RE	1시간 누적 강수유무(분)	TS*(종속변수)	1시간 평균 지면온도(℃)
TD	1시간 평균 이슬점 온도(℃)	WW	현천계 현천(S:눈/R:비/F:안개/H:박무/G:연무/C:맑음,X:모름)		
봄(106,999 rows x 14 columns) 여름(110,342 rows x 14 columns) 가을(110,328 rows x 14 columns) 겨울(110,374 rows x 14 columns)					
*결측값: -99, -99.9, -999					

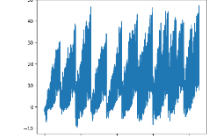
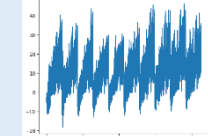
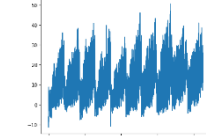
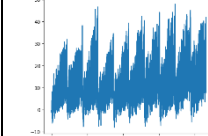
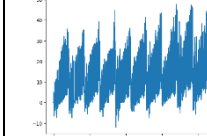
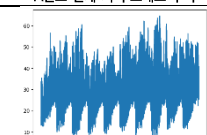
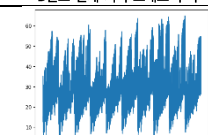
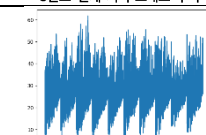
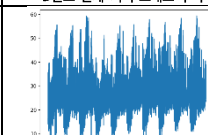
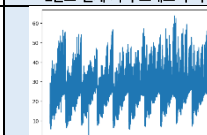
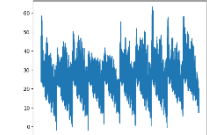
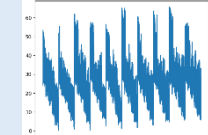
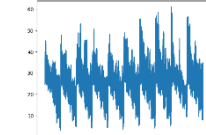
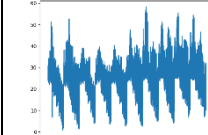
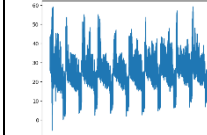
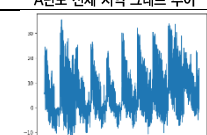
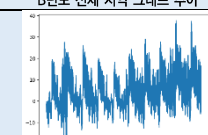
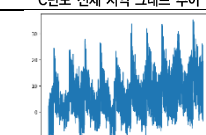
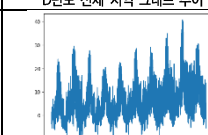
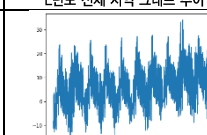
〈표 2〉 Test set 변수 개요

변수명	정의	변수명	정의	변수명	정의
STN	지점번호(a, b, c)	HM	1시간 평균 상대습도(%)	SI	1시간 누적 일사량(MJ)
YEAR	년도(F, G)	WS	1시간 평균 풍속(m/s)	SS	1시간 누적 일조량(초)
MMDDHH	월/일/시간	RN	1시간 누적 강수량(mm)	SN	00분에 측정된 적설 깊이(cm)
TA	1시간 평균 기온(℃)	RE	1시간 누적 강수유무(분)		
TD	1시간 평균 이슬점 온도(℃)	WW	현천계 현천(S:눈/R:비/F:안개/H:박무/G:연무/C:맑음,X:모름)		
봄(6,408 rows x 13 columns) 여름(6,624 rows x 13 columns) 가을(6,624 rows x 13 columns) 겨울(6,624 rows x 13 columns)					
*결측값: -99, -99.9, -999					

본 콘테스트에서 제공한 데이터는 A년 2월 ~ F년 1월까지의 국내 10개 지점의 11개의 기상 데이터를 가진 학습 데이터(Train set)와 F년 2월 ~ G년 1월까지의 특정 3지점의 11개의 기상 데이터를 가진 검증 데이터(Test set)이며 〈표 1〉, 〈표 2〉에서 해당 변수들을 확인할 수 있다. 본 콘테스트에서 제시한 계절별 기간은 봄(2월~4월), 여름(5월~7월), 가을(8월~10월), 겨울(11월~1월)이며 해당 기준을 바탕으로 학습 데이터와 검증 데이터를 계절별로 분리하여 사용하였다.

2.2. 계절별, 지역별, 연도별 지면온도 추이 파악

〈표 3〉 계절별, 지역별, 연도별 지면온도 파악

봄철 연도별 지역별 지면온도 추이				
A년도 전체 지역 그래프 추이	B년도 전체 지역 그래프 추이	C년도 전체 지역 그래프 추이	D년도 전체 지역 그래프 추이	E년도 전체 지역 그래프 추이
				
A년도 세부지역 지면온도 범위	B년도 세부지역 지면온도 범위	C년도 세부지역 지면온도 범위	D년도 세부지역 지면온도 범위	E년도 세부지역 지면온도 범위
1지역 (최저) -7.3℃ ~ (최대) 30.8℃ 2지역 (최저) -3.5℃ ~ (최대) 38.5℃ 3지역 (최저) -9.8℃ ~ (최대) 46.5℃ 4지역 (최저) -6.6℃ ~ (최대) 34.1℃ 5지역 (최저) -5.4℃ ~ (최대) 39.9℃ 6지역 (최저) -5.6℃ ~ (최대) 43.1℃ 7지역 (최저) -6.9℃ ~ (최대) 46.2℃ 8지역 (최저) -8.4℃ ~ (최대) 44.4℃ 9지역 (최저) -6.6℃ ~ (최대) 41.6℃ 10지역 (최저) -4.8℃ ~ (최대) 47.2℃	1지역 (최저) -11.5℃ ~ (최대) 40.4℃ 2지역 (최저) -18.3℃ ~ (최대) 36.2℃ 3지역 (최저) -12.4℃ ~ (최대) 43.7℃ 4지역 (최저) -13.1℃ ~ (최대) 30.0℃ 5지역 (최저) -10.2℃ ~ (최대) 30.6℃ 6지역 (최저) -12℃ ~ (최대) 36.5℃ 7지역 (최저) -11.2℃ ~ (최대) 45.7℃ 8지역 (최저) -9.4℃ ~ (최대) 43.1℃ 9지역 (최저) -9.8℃ ~ (최대) 46.3℃ 10지역 (최저) -8.8℃ ~ (최대) 41.4℃	1지역 (최저) -11.4℃ ~ (최대) 36.1℃ 2지역 (최저) -3.5℃ ~ (최대) 38.3℃ 3지역 (최저) -5.6℃ ~ (최대) 40.9℃ 4지역 (최저) -5.9℃ ~ (최대) 39.8℃ 5지역 (최저) -10.8℃ ~ (최대) 33.9℃ 6지역 (최저) -6.2℃ ~ (최대) 46.1℃ 7지역 (최저) -6℃ ~ (최대) 43.3℃ 8지역 (최저) -5.2℃ ~ (최대) 50.6℃ 9지역 (최저) -3℃ ~ (최대) 39.5℃ 10지역 (최저) -4.3℃ ~ (최대) 41.1℃	1지역 (최저) -6.1℃ ~ (최대) 32℃ 2지역 (최저) -4.2℃ ~ (최대) 38.3℃ 3지역 (최저) -7.7℃ ~ (최대) 46.7℃ 4지역 (최저) -7.5℃ ~ (최대) 34.7℃ 5지역 (최저) -8℃ ~ (최대) 39.4℃ 6지역 (최저) -6℃ ~ (최대) 42.5℃ 7지역 (최저) -7.5℃ ~ (최대) 44.3℃ 8지역 (최저) -7.2℃ ~ (최대) 48℃ 9지역 (최저) -4℃ ~ (최대) 45.2℃ 10지역 (최저) -5.6℃ ~ (최대) 42℃	1지역 (최저) -7.4℃ ~ (최대) 35.7℃ 2지역 (최저) -4.2℃ ~ (최대) 35.2℃ 3지역 (최저) -5.5℃ ~ (최대) 39.3℃ 4지역 (최저) -6.9℃ ~ (최대) 44.7℃ 5지역 (최저) -12.1℃ ~ (최대) 36.1℃ 6지역 (최저) -4.3℃ ~ (최대) 41.4℃ 7지역 (최저) -7.5℃ ~ (최대) 43.5℃ 8지역 (최저) -1.7℃ ~ (최대) 47.6℃ 9지역 (최저) -5.2℃ ~ (최대) 47.4℃ 10지역 (최저) -5.7℃ ~ (최대) 44.3℃
여름철 연도별 지역별 온도 추이				
A년도 전체 지역 그래프 추이	B년도 전체 지역 그래프 추이	C년도 전체 지역 그래프 추이	D년도 전체 지역 그래프 추이	E년도 전체 지역 그래프 추이
				
A년도 세부지역 지면온도 범위	B년도 세부지역 지면온도 범위	C년도 세부지역 지면온도 범위	D년도 세부지역 지면온도 범위	E년도 세부지역 지면온도 범위
1지역 (최저) 12.7℃ ~ (최대) 55.5℃ 2지역 (최저) 10℃ ~ (최대) 55.4℃ 3지역 (최저) 6.1℃ ~ (최대) 60.5℃ 4지역 (최저) 8.1℃ ~ (최대) 48.9℃ 5지역 (최저) 9.9℃ ~ (최대) 50.1℃ 6지역 (최저) 10.2℃ ~ (최대) 59.3℃ 7지역 (최저) 6.2℃ ~ (최대) 59.1℃ 8지역 (최저) 7.3℃ ~ (최대) 64.6℃ 9지역 (최저) 11.3℃ ~ (최대) 54.7℃ 10지역 (최저) 11.5℃ ~ (최대) 56.3℃	1지역 (최저) 4℃ ~ (최대) 57.3℃ 2지역 (최저) 5.5℃ ~ (최대) 53.6℃ 3지역 (최저) 5.4℃ ~ (최대) 60.4℃ 4지역 (최저) 6.7℃ ~ (최대) 57.8℃ 5지역 (최저) 7.2℃ ~ (최대) 57.1℃ 6지역 (최저) 6.9℃ ~ (최대) 53.6℃ 7지역 (최저) 5.7℃ ~ (최대) 59.1℃ 8지역 (최저) 3.6℃ ~ (최대) 64.4℃ 9지역 (최저) 7.1℃ ~ (최대) 64.9℃ 10지역 (최저) 7.5℃ ~ (최대) 64.7℃	1지역 (최저) 6.2℃ ~ (최대) 53.5℃ 2지역 (최저) 6.1℃ ~ (최대) 51.6℃ 3지역 (최저) 5.7℃ ~ (최대) 61.8℃ 4지역 (최저) 5.4℃ ~ (최대) 51.9℃ 5지역 (최저) 7.4℃ ~ (최대) 51.9℃ 6지역 (최저) 5℃ ~ (최대) 55.4℃ 7지역 (최저) 7.6℃ ~ (최대) 55.5℃ 8지역 (최저) 5.3℃ ~ (최대) 52.4℃ 9지역 (최저) 11.3℃ ~ (최대) 53.3℃ 10지역 (최저) 8.6℃ ~ (최대) 52.8℃	1지역 (최저) 8.3℃ ~ (최대) 55.5℃ 2지역 (최저) 8.6℃ ~ (최대) 55.3℃ 3지역 (최저) 7.2℃ ~ (최대) 59.2℃ 4지역 (최저) 8.2℃ ~ (최대) 51.2℃ 5지역 (최저) 7.4℃ ~ (최대) 55.3℃ 6지역 (최저) 10℃ ~ (최대) 55.6℃ 7지역 (최저) 8.4℃ ~ (최대) 57.8℃ 8지역 (최저) 8.6℃ ~ (최대) 56.6℃ 9지역 (최저) 11.1℃ ~ (최대) 58.3℃ 10지역 (최저) 10.9℃ ~ (최대) 59.3℃	1지역 (최저) 5.7℃ ~ (최대) 55.5℃ 2지역 (최저) 6.3℃ ~ (최대) 54.3℃ 3지역 (최저) 0℃ ~ (최대) 47.1℃ 4지역 (최저) 7.6℃ ~ (최대) 56.7℃ 5지역 (최저) 6.3℃ ~ (최대) 49.2℃ 6지역 (최저) 5.6℃ ~ (최대) 57.4℃ 7지역 (최저) 7.1℃ ~ (최대) 57.8℃ 8지역 (최저) 7℃ ~ (최대) 63.9℃ 9지역 (최저) 8.5℃ ~ (최대) 59.4℃ 10지역 (최저) 7.2℃ ~ (최대) 56.1℃
가을철 연도별 지역별 온도 추이				
A년도 전체 지역 그래프 추이	B년도 전체 지역 그래프 추이	C년도 전체 지역 그래프 추이	D년도 전체 지역 그래프 추이	E년도 전체 지역 그래프 추이
				
A년도 세부지역 지면온도 범위	B년도 세부지역 지면온도 범위	C년도 세부지역 지면온도 범위	D년도 세부지역 지면온도 범위	E년도 세부지역 지면온도 범위
1지역 (최저) -1.9℃ ~ (최대) 58.4℃ 2지역 (최저) -0.2℃ ~ (최대) 44.9℃ 3지역 (최저) -2℃ ~ (최대) 50.3℃ 4지역 (최저) 1.2℃ ~ (최대) 39.8℃ 5지역 (최저) 1.8℃ ~ (최대) 41.6℃ 6지역 (최저) 0.6℃ ~ (최대) 55.4℃ 7지역 (최저) 0.4℃ ~ (최대) 51.8℃ 8지역 (최저) 9.7℃ ~ (최대) 63.3℃ 9지역 (최저) 5.6℃ ~ (최대) 57.8℃ 10지역 (최저) 7.3℃ ~ (최대) 58.2℃	1지역 (최저) 0.2℃ ~ (최대) 53.3℃ 2지역 (최저) 0.8℃ ~ (최대) 55.4℃ 3지역 (최저) 0.3℃ ~ (최대) 61.8℃ 4지역 (최저) 2.5℃ ~ (최대) 57.4℃ 5지역 (최저) 1.2℃ ~ (최대) 56.4℃ 6지역 (최저) 1.6℃ ~ (최대) 65.1℃ 7지역 (최저) 2.9℃ ~ (최대) 60.5℃ 8지역 (최저) 4.6℃ ~ (최대) 63.5℃ 9지역 (최저) 5.6℃ ~ (최대) 66.7℃ 10지역 (최저) 5.6℃ ~ (최대) 57.2℃	1지역 (최저) 3℃ ~ (최대) 45.2℃ 2지역 (최저) 4.6℃ ~ (최대) 47.8℃ 3지역 (최저) 4.1℃ ~ (최대) 53.3℃ 4지역 (최저) 4.4℃ ~ (최대) 50.7℃ 5지역 (최저) 3.6℃ ~ (최대) 48.2℃ 6지역 (최저) 5.2℃ ~ (최대) 49.1℃ 7지역 (최저) 4.1℃ ~ (최대) 55.5℃ 8지역 (최저) 4.3℃ ~ (최대) 58.6℃ 9지역 (최저) 9.4℃ ~ (최대) 61.2℃ 10지역 (최저) 7.8℃ ~ (최대) 56.6℃	1지역 (최저) 0.6℃ ~ (최대) 51.3℃ 2지역 (최저) 1.3℃ ~ (최대) 52.7℃ 3지역 (최저) 4℃ ~ (최대) 39.1℃ 4지역 (최저) 3.4℃ ~ (최대) 39.6℃ 5지역 (최저) 2.6℃ ~ (최대) 40.3℃ 6지역 (최저) 2.5℃ ~ (최대) 47.6℃ 7지역 (최저) 4.8℃ ~ (최대) 58.3℃ 8지역 (최저) 4℃ ~ (최대) 51.3℃ 9지역 (최저) 6.7℃ ~ (최대) 55.4℃ 10지역 (최저) 6.9℃ ~ (최대) 54.4℃	1지역 (최저) -5.5℃ ~ (최대) 59.1℃ 2지역 (최저) 3.3℃ ~ (최대) 53.6℃ 3지역 (최저) 2.4℃ ~ (최대) 55.1℃ 4지역 (최저) 3.4℃ ~ (최대) 55.1℃ 5지역 (최저) 2℃ ~ (최대) 47℃ 6지역 (최저) 2.6℃ ~ (최대) 47.4℃ 7지역 (최저) 4.6℃ ~ (최대) 53.7℃ 8지역 (최저) 6.3℃ ~ (최대) 47.9℃ 9지역 (최저) 6.6℃ ~ (최대) 57.1℃ 10지역 (최저) 6.8℃ ~ (최대) 59.2℃
겨울철 연도별 지역별 온도 추이				
A년도 전체 지역 그래프 추이	B년도 전체 지역 그래프 추이	C년도 전체 지역 그래프 추이	D년도 전체 지역 그래프 추이	E년도 전체 지역 그래프 추이
				
A년도 세부지역 지면온도 범위	B년도 세부지역 지면온도 범위	C년도 세부지역 지면온도 범위	D년도 세부지역 지면온도 범위	E년도 세부지역 지면온도 범위
1지역 (최저) -10.4℃ ~ (최대) 24.5℃ 2지역 (최저) -14.5℃ ~ (최대) 35.5℃ 3지역 (최저) -10.2℃ ~ (최대) 25.3℃ 4지역 (최저) -8.4℃ ~ (최대) 26.5℃ 5지역 (최저) -8.3℃ ~ (최대) 22.8℃ 6지역 (최저) -8.3℃ ~ (최대) 30.2℃ 7지역 (최저) -10.7℃ ~ (최대) 32.1℃ 8지역 (최저) -10.2℃ ~ (최대) 30.5℃ 9지역 (최저) -7.4℃ ~ (최대) 33.9℃ 10지역 (최저) -5℃ ~ (최대) 24.1℃	1지역 (최저) -12.6℃ ~ (최대) 19.3℃ 2지역 (최저) -19.9℃ ~ (최대) 27.5℃ 3지역 (최저) -13.6℃ ~ (최대) 26℃ 4지역 (최저) -12.8℃ ~ (최대) 20℃ 5지역 (최저) -10.5℃ ~ (최대) 18.1℃ 6지역 (최저) -12.1℃ ~ (최대) 24.2℃ 7지역 (최저) -12.8℃ ~ (최대) 29℃ 8지역 (최저) -12.4℃ ~ (최대) 27.7℃ 9지역 (최저) -7.9℃ ~ (최대) 37.7℃ 10지역 (최저) -9.6℃ ~ (최대) 37.3℃	1지역 (최저) -12.1℃ ~ (최대) 24.4℃ 2지역 (최저) -7.3℃ ~ (최대) 23.8℃ 3지역 (최저) -11.1℃ ~ (최대) 27.9℃ 4지역 (최저) -11.2℃ ~ (최대) 26.5℃ 5지역 (최저) -11.8℃ ~ (최대) 23.3℃ 6지역 (최저) -9.2℃ ~ (최대) 33.7℃ 7지역 (최저) -9.4℃ ~ (최대) 31.9℃ 8지역 (최저) -7.4℃ ~ (최대) 33.6℃ 9지역 (최저) -4.8℃ ~ (최대) 30.8℃ 10지역 (최저) -6.3℃ ~ (최대) 35.3℃	1지역 (최저) -11.2℃ ~ (최대) 24℃ 2지역 (최저) -11.9℃ ~ (최대) 29.5℃ 3지역 (최저) -8℃ ~ (최대) 27.6℃ 4지역 (최저) -11.7℃ ~ (최대) 20.5℃ 5지역 (최저) -8.1℃ ~ (최대) 22.5℃ 6지역 (최저) -8.1℃ ~ (최대) 28.5℃ 7지역 (최저) -9.1℃ ~ (최대) 29℃ 8지역 (최저) -5.5℃ ~ (최대) 35℃ 9지역 (최저) -4.8℃ ~ (최대) 40.7℃ 10지역 (최저) -5.4℃ ~ (최대) 30.6℃	1지역 (최저) -12.8℃ ~ (최대) 23.8℃ 2지역 (최저) -17.5℃ ~ (최대) 23.3℃ 3지역 (최저) -12.8℃ ~ (최대) 25.6℃ 4지역 (최저) -13.9℃ ~ (최대) 23.6℃ 5지역 (최저) -9.3℃ ~ (최대) 26.6℃ 6지역 (최저) -12.1℃ ~ (최대) 27.2℃ 7지역 (최저) -11.8℃ ~ (최대) 27.7℃ 8지역 (최저) -3.9℃ ~ (최대) 27.1℃ 9지역 (최저) -6.4℃ ~ (최대) 34.1℃ 10지역 (최저) -8.8℃ ~ (최대) 27.5℃

위 <표 3>과 같이 계절별, 연도별로 전체 지면온도 수치를 계산 및 시각화 했을 때, 전반적인 지면온도 범위의 폭이 넓은 것을 확인할 수 있었다. 특히, **봄의 지면온도 범위는 B년도에 $-18.3^{\circ}\text{C} \sim 46.3^{\circ}\text{C}$ 로 폭이 가장 넓었으며, 여름은 E년도에 $0^{\circ}\text{C} \sim 63.8^{\circ}\text{C}$ 의 넓은 지면온도 범위를 보였다. 가을 및 겨울은 B년도에 각각 $0.2^{\circ}\text{C} \sim 65.7^{\circ}\text{C}$, $-19.9^{\circ}\text{C} \sim 37.7^{\circ}\text{C}$ 로 가장 넓은 지면온도 범위폭을 나타냈다. 이러한 결과는 봄/여름/가을/겨울의 범위가 통상적인 구간과는 다를 뿐만 아니라, 10개 지역별 기상 정보가 큰 차이를 보이기 때문으로 추정할 수 있다.**

3. 데이터 전처리 및 파생변수 생성 과정

<표 4> 계절별 데이터 결측치 개수 및 비율

봄				여름				가을				겨울			
변수명	관측된 결측값	개수(개)	비율(%)	변수명	관측된 결측값	개수(개)	비율(%)	변수명	관측된 결측값	개수(개)	비율(%)	변수명	관측된 결측값	개수(개)	비율(%)
RE	[~99]	301	0.0028	RE	[~99]	736	0.0067	RE	[~99]	481	0.0044	RE	[~99]	424	0.0038
TA	[~99.9]	111	0.001	TA	[~99.9]	99	0.0009	TA	[~99.9]	117	0.0011	TA	[~99.9]	106	0.001
TD	[~99.9]	112	0.001	TD	[~99.9]	106	0.001	TD	[~99.9]	120	0.0011	TD	[~99.9]	111	0.001
HM	[~99.9]	103	0.001	HM	[~99.9]	92	0.0008	HM	[~99.9]	95	0.0009	HM	[~99.9]	91	0.0008
WS	[~99.9]	167	0.0016	WS	[~99.9]	162	0.0015	WS	[~99.9]	256	0.0023	WS	[~99.9]	225	0.002
RN	[~99.9]	2,277	0.0213	RN	[~99.9]	1,955	0.0177	RN	[~99.9]	1,291	0.0117	RN	[~99.9]	2,275	0.0206
TS	[~99.9]	114	0.0011	TS	[~99.9]	111	0.001	TS	[~99.9]	111	0.001	TS	[~99.9]	100	0.0009
SI	[~99.9]	48,956	0.4575	SI	[~99.9]	41,873	0.3795	SI	[~99.9]	49,451	0.4482	SI	[~99.9]	59,549	0.5395
SS	[~99.9]	49,045	0.4584	SS	[~99.9]	41,520	0.3763	SS	[~99.9]	48,775	0.4421	SS	[~99.9]	59,526	0.5393
SN	[~99.9]	104,778	0.9792	SN	[~99.9]	110,342	1.0	SN	[~99.9]	110,328	1.0	SN	[~99.9]	105,401	0.9549

본 데이터의 계절별 각 변수의 결측치 개수 및 비율은 <표 4>와 같다. 전체적인 결측값의 분포는 각 계절별로 유사한 양상을 보였고, 특히 SI, SS, SN의 결측치 비율이 높게 나온 것으로 확인되었다. 또한 타 기상 변수 정보와 맞지 않는 현천(WW) 데이터가 존재하는 것을 파악할 수 있었다. 따라서 기상청에서 명시한 기간별로 전체 데이터셋을 나누어 각 계절별 데이터셋을 생성한 다음 아래와 같은 세부적인 분석을 진행하여 결측치를 처리하였다.

3.1. 결측치 행 제거

우선 계절별 데이터셋을 분할하기 전, 전체 훈련 데이터 기준으로 종속변수를 포함한 모든 변수 값이 결측치인 197개 행을 제거하였다. 또한 계절별 데이터셋으로 분할한 다음 종속 변수인 지면온도(TS)의 값이 결측치인 행을 제거하였다. 결과적으로 봄, 여름, 가을, 겨울 데이터셋 순으로 각각 114개, 111개, 111개, 100개의 총 436개 행을 제거하였다.

3.2. 일사량(SI), 일조량(SS) 변수 결측치 처리

일사량(SI), 일조량(SS) 변수는 각각 1시간 누적 일사량 및 1시간 누적 일조량이므로 해가 떠있는 시간만 관측이 가능하다. 그렇기 때문에 일출부터 일몰까지만 해당 변수들의 관측이 가능하다. 이에 따라 각 월 별로 일사량(SI), 일조량(SS) 데이터를 확인했을 때, 특정 시간대의 구간에서 결측치가 몰려있는 양상을 확인하였다. 따라서 해당 구간의 결측치를 0으로 대체하였다.

3.3. 적설(SN) 변수 결측치 처리

봄 데이터는 2월 17일 이후 적설(SN) 변수가 모두 결측치로 파악되어 2월 17일 이후에는 적설(SN)이 없는 것으로 처리했다. 이를 위해 2월 17일 이후까지의 적설(SN) 결측치는 0으로 처리하였다. **여름 및 가을 데이터**에서는 적설(SN) 변수가 모두 결측치로 관측되었다. 따라서 해당 계절의 결측치는 모두 0으로 처리하였다. **겨울 데이터**에서는 적설이 처음 관측된 날짜가 11월 24일이었다. 따라서 겨울 데이터의 11월 24일 이전의 적설(SN) 결측치는 0으로 처리하였다.

3.4. 현천(WW) 이상치 데이터 제거

데이터 EDA를 통해 지면 온도를 파악하는 과정에서 현천(WW)이 눈(S)인 경우, 기온(TA) 및 지면 온도(TS)가 10°C 를 넘어서는 이상치 데이터를 발견하여 <표 5>와 같은 과정을 거친 후, 봄 데이터에서 14개, 여름 데이터에서 11개, 가을 데이터에서 11개의 총 36개의 이상치 데이터를 제거하였다. 추가적으로 기온이 10°C 내에서는 이슬점 온도 및 상대습도, 풍속의 영향에 따라 눈이 내릴 수 있으며, 본 분석에서 10°C 내의 눈(S)으로 기록된 변수들의 이슬점 온도, 상대습도, 풍속의 수치를 모두 확인 후 사용하였다.

〈표 5〉 현천 이상치 데이터 제거 방식



3.5. 다항식 보간법(Polynomial Interpolation)

선형 보간법은 관측치를 직선으로 연결하여 그 사이의 결측치를 해당 직선 위의 값으로 추측하는 방법이다. 하지만 단순히 선형으로 보간할 경우 다차원의 변수로 이루어진 본 데이터 셋에서는 해당 보간법의 정확도가 현저히 낮아질 수 있다. 이에 대한 대안으로 실제 연구 중 기상 데이터에, 직선이 아닌 다항식을 활용한 보간법이 많이 사용되는 것을 확인하였다 (Chin et al., 2023; Antal et al., 2021). 본 실험에서는 2차 다항식을 활용하는 보간법을 사용하였으며, 변수에 따라 해당 결측치의 맨 앞 또는 뒤에 관측치가 존재하지 않아 다항식 보간법이 이뤄지지 않은 경우에는 해당 결측치의 앞 또는 뒤의 관측치 값으로 대체하였다. 최종적으로 보간한 결과 종속변수인 지면온도(TS)외의 결측치가 존재하던 10개의 변수에서 결측치가 모두 대체되었다.

3.6. 파생변수 생성

〈표 6〉 파생변수 소개

생성변수	생성 방법	활용된 계절별 모델
월(month), 일(day), 시간(hour)	월일시간(mmddhh) 변수로부터 월(month), 일(day), 시간(hour)변수 생성	봄, 여름, 가을, 겨울
강수여부(rs_yn)	현천(ww) 변수가 눈(S)이거나 비(R)인 경우는 1, 그렇지 않은 경우 0으로 처리	봄, 여름, 가을, 겨울
체감온도(sense_ta)	$13.12 + 0.6215T - 11.37W^{0.16} + 0.3965W^{0.16}T$ *(T: 기온(ta), W: 풍속(ws))	봄, 겨울
화씨온도(f_ta)	$(T \times 9/5) + 32$ *(T: 기온(ta))	봄, 여름, 가을, 겨울
열지수(h_idx)	Rothfusz의 회귀방정식으로 화씨 80도 미만의 조건에서는 유효하지 않음 따라서 화씨온도가 80도 이상이면 아래의 식을 적용하고 그렇지 않은 경우 해당 기온을 열지수로 대체함 (1) $value = -42.379 + 2.04901523F + 10.14333127H - 0.22475541FH - 0.00683783F^2 - 0.05481717H^2 + 0.00122874F^2H + 0.00085282FH^2 - 0.00000199F^2H^2$ (2) $value2 = (value - 32) \times 5/9$ (3) $value3 = round(value2)/10.0$ *(F: 화씨온도(f_ta), H: 상대습도(hm))	여름, 가을
불쾌지수(u_idx)	$(T - 0.55 \times (1 - (0.01 \times H)) \times (T - 14.5)$ *(T: 기온(ta), H: 상대습도(hm))	여름, 가을

본 실험을 진행하기에 앞서 내부 파생변수들을 〈표 6〉과 같이 생성하였다. 월(month), 일(day), 시간(hour) 변수는 앞선 결측치 처리과정에서 월, 일, 시간 단위로 세부적인 판단이 필요한 상황이 있어 파생변수로 생성하였다. 강수여부의 경우 현천(ww)의 변수를 활용하여 해당 파생변수를 생성하였다. 체감온도는 인간이 느끼는 더위나 추위를 수량적으로 나타낸 것으로 여름철(5~9월)과 겨울철(10~익년 4월)을 구분하여 계산이 가능하다. 하지만 본 공모전 데이터의 경우 여름철 체감온도를 산출하는 값 중 습구온도가 따로 존재하지 않기 때문에, 겨울철 체감온도만 활용하여 봄과 겨울철의 데이터 셋에서 체감온도를 산출하였다. 화씨온도의 경우 열지수를 산출하기 위해 우선적으로 계산하여 생성을 하였다. 열지수는 기온과 습도에 따라 사람이 실제로 느끼는 더위를 지수화한 것이며, 지면 온도(TS)가 30~50°C의 값을 전반적으로 나타내는 여름과 가을 데이터 셋에서 열지수 파생변수를 생성하였다. 불쾌지수는 사람이 느끼는 불쾌감의 정도이며 기온과 습도의 조합으로 값을 산출할 수 있다. 불쾌지수는 열지수와 마찬가지로 30~50°C의 값들을 나타내는 여름과 가을 데이터 셋에서 해당 파생변수를 생성하였다.

4. 기본 모델 구축

모든 전처리를 완료한 최종 데이터를 활용할 때, 본 실험에서는 이상치를 모두 반영하여 모델을 구축하기 때문에 중앙값과 IQR을 사용하여 이상치의 영향을 최소화해주는 RobustScaler를 사용하여 데이터를 표준화하였다(Raju et al., 2020). 이후 기상 예측과 관련된 선행연구들을 조사해봤을 때, LGBM(Light Gradient Boosting Model), CatBoost, XGBoost와 같은 boosting 계열 모델들이 좋은 성능을 보인 것을 확인할 수 있었다(Han et al., 2023; Niu et al., 2021; Ma & Ji, 2020). 따라서 기본 모델을 LGBM, CatBoost, XGBoost로 선정하였다. 기본 모델의 random_state 값은 42로 고정하였으며, Parameter 값은 기본으로 설정하여 각 모델에 데이터를 학습시키고 MAE를 산출하였다. MAE 산출 결과, 봄, 여름 및 겨울 데이터에서는 LGBM 모델이, 가을 데이터에서는 CatBoost 모델이 각각 기본 모델 중에서 가장 좋은 성능을 보였다.

4.1 변수 중요도 파악

앞선 기본 모델을 구축한 결과 중 계절별로 가장 잘 나온 모델을 기준으로 내장 함수(feature_importances_)를 적용해 변수 중요도가 0.1%(LGBM = 0.001, CatBoost = 0.1) 미만인 변수를 각 계절별 모델에서 제거하였으며, 그 결과는 <표 7>에서 찾아볼 수 있다. *(단, 겨울 모델 변수 중 월(month)의 경우 모든 모델의 변수 중요도를 종합적으로 파악해 본 결과, 다른 모델에서 모두 0의 값을 나타내서 해당 모델에서는 제거하기로 결정하였음)

<표 7> 각 계절별 모델 변수 중요도 파악

변수명 / 계절	봄(LGBM)	여름(LGBM)	가을(CatBoost)	겨울(LGBM)
월일시간(mmddhh)	0.1466	0.1736	1.841757	0.1286
기온(ta)	0.1326	0.1173	20.172713	0.131
이슬점 온도(td)	0.0923	0.0873	1.493564	0.0666
상대습도(hm)	0.079	0.0833	2.121682	0.0656
풍속(ws)	0.0666	0.07	1.106124	0.056
강수(rn)	0.01	0.0043	0.055587(제거)	0.0053
강수유무(re)	0.0113	0.0083	0.174311	0.0123
현전(ww)	0.0256	0.0356	0.43827	0.025
일사량(si)	0.1536	0.136	27.72064	0.142
일조시간(ss)	0.0323	0.036	2.354042	0.026
적설(sn)	0.0443	0(제거)	0(제거)	0.1836
월(month)	0(제거)	0(제거)	0.053896(제거)	*0.0036(제거)
일(day)	0.0456	0.0553	1.226451	0.038
시간(hour)	0.1113	0.112	5.25577	0.0766
강수여부(rs_yn)	0.0046	0.003	0.105852	0.0006(제거)
체감온도(sense_ta)	0.0436	해당 없음	해당 없음	0.0386
화씨온도(f_ta)	0(제거)	0(제거)	19.337002	0(제거)
열지수(h_idx)	해당 없음	0.023	10.849456	해당 없음
불쾌지수(u_idx)	해당 없음	0.0543	4.35758	해당 없음

*LGBM 모델 = 변수 중요도 총 합 1 / CatBoost 모델 = 변수 중요도 총 합 100

4.2. 모델 튜닝 과정

1) 변수 중요도를 파악해 Cut-Off 기준 0.1% 미만인 변수들을 제거한 Baseline 모델들의 각각의 성능을 높이기 위해 최적의 Hyper Parameter를 자동으로 찾아주는 Optuna 소프트웨어를 적용하여 결과를 산출하였는데, 해당 작업 관련해서는 A, B, C, D년도를 학습 데이터, 나머지 E년도를 검증 데이터로 지정하여 Optuna를 적용 후 최적의 값을 산출하였다. 그 결과 LGBM의 파라미터 튜닝이 모든 계절별 모델에서 가장 높게 나왔으며 Hyper Parameter 값은 아래 <표 8>과 같다.

<표 8> Optuna Hyper Parameter 적용 후 모델 성능 결과

봄		여름		가을		겨울	
Model	MAE	Model	MAE	Model	MAE	Model	MAE
LGBM(Best)	2.724	LGBM(Best)	3.232	LGBM(Best)	2.061	LGBM(Best)	1.954
CatBoost	3.003	CatBoost	3.774	CatBoost	2.275	CatBoost	2.232
XGBoost	3.656	XGBoost	3.857	XGBoost	2.302	XGBoost	2.235
Best Model Hyper Parameter 값		Best Model Hyper Parameter 값		Best Model Hyper Parameter 값		Best Model Hyper Parameter 값	

N_estimators = 265 Max_depth = 6 Learning_rate = 0.03848576560913784 Min_child_samples = 24 Colsample_bytree = 0.9275432756814064 Subsample = 0.4818277531049522	N_estimators = 1380 Max_depth = 7 Learning_rate = 0.00379864792169426 Min_child_samples = 19 Colsample_bytree = 0.47828626887767617 Subsample = 0.4474447500029784	N_estimators = 1968 Max_depth = 15 Learning_rate = 0.14199190766877956 Min_child_samples = 26 Colsample_bytree = 0.9471423809959226 Subsample = 0.8804923171057708	N_estimators = 2000 Max_depth = 15 Learning_rate = 0.19481757590060347 Min_child_samples = 7 Colsample_bytree = 0.9898081343609673 Subsample = 0.4668480569567963
---	---	---	--

2) <표 8>의 성능 결과 확인 후, 성능의 추가적인 개선을 위해 Ensemble 기법을 적용하였다. Ensemble을 구축할 때, 기본 LGBM, CatBoost, XGBoost를 전부 넣어 모델을 구축하였다. 아래 <표 9>의 Ensemble 모델 성능의 결과를 확인했을 때, Hyper Parameter를 적용한 각 계절별 단일 LGBM 모델들에 비해 모두 상향된 결과가 나온 것을 확인할 수 있었다. 추가적인 실험으로 Ensemble 모델 성능 향상을 위해 Ensemble 내 각 모델에 Optuna를 활용하여 조정된 Hyper Parameter를 투입하여 학습하였으나, 기본 모델에 비해 성능이 떨어지는 것으로 나타났다. 따라서 기본 LGBM, CatBoost, XGBoost 모델을 포함하는 Ensemble을 최종 모델로 선정하고, 추가적으로 Ensemble 내 각 모델의 예측 가중치를 지속적으로 조정하여 아래와 같은 최종 MAE 결과를 산출하였다.

<표 9> Ensemble 모델 성능 결과

봄 Ensemble	여름 Ensemble	가을 Ensemble	겨울 Ensemble
[기본 Ensemble] MAE = 1.789	[기본 Ensemble] MAE = 2.016	[기본 Ensemble] MAE = 1.926	[기본 Ensemble] MAE = 1.982
LGBM*0.4 CatBoost*0.4 XGBoost*0.2	LGBM*0.3 CatBoost*0.6 XGBoost*0.1	LGBM*0.7 CatBoost*0.2 XGBoost*0.1	LGBM*0.6 CatBoost*0.3 XGBoost*0.1
[가중치 적용 Ensemble] MAE = 1.784	[가중치 적용 Ensemble] MAE = 2.013	[가중치 적용 Ensemble] MAE = 1.71	[가중치 적용 Ensemble] MAE = 1.879

최종 MAE 평균 값 = 1.846

*기본 Ensemble(기본 LGBM+기본 CatBoost+기본 XGBoost)

5. 서비스 활용방안 및 기대효과

지면온도는 농업, 에너지, 건축, 환경, 국방 등 다양한 분야에서 적극적으로 활용되고 있으며, 특히 지면온도를 활용한 자연재해 예방과 건축 및 도시계획 수립에 있어 대단히 중요한 역할을 한다. 구체적으로 지면온도를 정확히 예측하는 것은 미래에 발생할 수 있는 자연재해로 인한 피해를 사전에 예방할 수 있으며, 도시계획 정책 수립 및 건축물 건설 등에 지면온도 정보를 활용하여 기후 친화적인 도시를 개발하여 시민들에게 편의를 제공할 수 있다. 본 공모전에서 제공받은 데이터를 통해 구축한 계절별 통합 모델은 효과적으로 지면온도를 예측하여 앞서 언급한 분야에서 활용이 가능할 것으로 기대된다.

추가적으로, 제안된 모델과 분석 방안에 다양한 데이터 및 모델을 추가하여 보다 정확하고 종합적인 평가를 제공할 수 있는 통합 모델을 개발할 수도 있다. 최근 전국적으로 기온상승, 집중 호우 및 가뭄과 태풍의 증가, 해수면 상승 등 실생활에 큰 피해를 주는 이상 기후로 인한 자연재해가 더욱 빈번하게 나타나고 있어 사회적으로 큰 문제를 야기하고 있다. 본 공모전에서 구축한 지면온도 예측 모델에 다양한 기후 관련 데이터 및 모델을 결합한다면 이러한 이상 기후에 대한 정확한 예측과 동시에 종합적인 평가 및 진단이 가능할 수 있다. 예를 들어, 대기 예측 모델과 지면온도 예측 모델을 결합하여 구축한 모델을 통해 급격한 기후 변화에 대해 정확히 예측한다면, 사전에 적절한 예방 조치를 취하여 피해를 최소화할 수 있다.

본 실험의 EDA 분석과정에서 계절별, 지역별, 연도별로 지면온도 세부 추세를 살펴봤을 때, 지역별로 지면온도 분포가 광범위한 것을 확인하였다. 제안된 모델은 모든 지역을 고려해서 계절별로 모델을 통합 구축하였으나, 실무적으로 더욱 정확한 지면온도를 예측하기 위해서는 각 지역의 특성이 반영된 세분화된 모델의 구축이 필요하다. 따라서 추후 각 지역의 계절별로 모델을 세부적으로 분할하여 데이터 탐색과정을 거친 계절별-지역별 모델을 구축한다면 더욱 실효적인 서비스가 될 것으로 기대한다.

참고문헌

- Antal, A., Guerreiro, P. M., & Cheval, S. (2021). Comparison of spatial interpolation methods for estimating the precipitation distribution in Portugal. *Theoretical and Applied Climatology*, 145(3-4), 1193-1206.
- Chin, R. J., Lai, S. H., Loh, W. S., Ling, L., & Soo, E. Z. X. (2023). Assessment of Inverse Distance Weighting and Local Polynomial Interpolation for Annual Rainfall: A Case Study in Peninsular Malaysia. *Engineering Proceedings*, 38(1), 61.
- Han, W., Duan, S. B., Tian, H., & Lian, Y. (2023). Estimation of land surface temperature from AMSR2 microwave brightness temperature using machine learning methods. *International Journal of Remote Sensing*, 1-22.
- Ma, X., Fang, C., & Ji, J. (2020). Prediction of outdoor air temperature and humidity using Xgboost. In *IOP conference series: earth and environmental science* (Vol. 427, No. 1, p. 012013). IOP Publishing.
- Niu, D., Diao, L., Zang, Z., Che, H., Zhang, T., & Chen, X. (2021). A machine-learning approach combining wavelet packet denoising with Catboost for weather forecasting. *Atmosphere*, 12(12), 1618.
- Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020, August). Study the influence of normalization/transformation process on the accuracy of supervised classification. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 729-735). IEEE.