Heat? Hit!

(히트힛)

일일 열사병 환자 수 예측 프로그램

(Daily Heatstroke Patient Prediction Program)

<3 조>

발표자: 김인영

팀원: 김인영, 이성준, 정유진, 최미선

Github: https://github.com/Kiminyoung3/240821_Project

목차

1. 열사병 관련 머신러닝 모델 개발 관련 요약

본 프로젝트는 건설현장에서 발생하는 열사병 환자 수를 예측하기 위한 머신러닝 모델을 개발하는 것을 목표로 한다. 최근 여름철 기온 상승과 작업 강도 증가로 인해 열사병 발생이 급증하고 있으며, 2022 년과 2023 년 여름에는 기록적인 폭염으로 건설업에서 열사병 환자 수와 인명 피해가 크게 증가한 바 있다.

이러한 문제를 해결하고자 기상 데이터와 작업 환경 데이터를 활용한 예측 모델을 구축하였다. 주요 독립변수로는 월별 최고기온, 평균기온, 평균습도, 강수량, 일교차, 체감온도, 불쾌지수, 전날 이송된 환자 수, 5 일간 평균 열사병 이송 환자 수 등을 사용하였다. 이를 통해 위험도가 높은 작업 조건에서 열사병 발생 가능성을 예측할 수 있는 모델을 설계하였다.

모델링에는 Gradient Boosting 학습 모델을 적용하였으며, K-Fold 교차 검증을 통해 예측 성능을 평가하였다. 이 모델은 다수의 독립변수를 입력하여 하루 단위의 열사병 이송 환자 수를 예측하고, 관리자에게 적절한 경고와 대응책을 제공할 수 있도록 설계되었다. 이를 통해 작업 일정을 조정하거나 냉방 시설을 설치하는 등의 사전 조치를 통해 열사병 발생을 효과적으로 줄이는 데 기여할 수 있을 것으로 기대된다.

2. 개발 목적

1) 머신러닝 모델 활용 대상

① 현장 작업자 안전 관리

가) 예측 기반 예방 조치

a. 상황 인식 향상: 머신러닝 모델을 통해 기온, 습도 등 날씨 정보를 기반으로 열사병 발생 가능성을 예측함으로써 작업자들에게 필요한 안전 조치를 사전에 안내할 수 있도록 한다.

b. 실시간 경고 시스템: 날씨 데이터와 예측 모델을 활용하여 현장 작업자들에게 실시간으로 열사병 위험 정보를 제공하며, 이를 통해 위험한 조건에서는 작업을 연기하거나 중단할 수 있도록 한다.

나) 작업 일정 조정

a. 작업 시간 조정: 열사병 위험도가 높은 날씨 조건을 예측하여, 고온의 날에는 작업 일정을 조정하거나 작업 시간을 변경하여 작업자의 건강을 보호한다.

b. 휴식 주기 설정: 예측된 열사병 발생 가능성을 고려하여 적절한 휴식 주기를 설정하고, 충분한 수분 섭취와 휴식을 권장한다.

② 건강 관리 및 정책 결정

가) 정책 및 지침 개발

a. 보건 정책 수립: 예측 모델을 기반으로 열사병 발생 패턴을 분석하여, 지역적 및 산업별 보건 정책과 지침을 개발한다.

b. 예방 교육 자료 제공: 열사병 발생 위험이 높은 상황에서의 예방 교육 자료와 가이드를 제공하여, 작업자들이 자가 보호 및 응급 조치를 적절히 수행할 수 있도록 한다.

나) 의료 대응 계획

a. 응급 대응 시스템: 열사병 환자 수 예측 결과를 바탕으로 응급 의료 대응 시스템을 구축하고, 인근 병원이나 응급 센터의 자원을 효율적으로 배분한다.

b. 건강 모니터링 시스템: 작업자들의 건강 상태를 모니터링하고, 예측된 열사병 발생 가능성에 따라 맞춤형 건강 관리 서비스를 제공하는 시스템을 구축한다.

2) 개발의 의의

① 작업자의 건강 및 안전 강화

가) 예방 조치: 예측 프로그램을 통해 열사병 발생 가능성을 사전에 파악하고, 작업자에게 적절한 예방 조치를 안내하여 건강 문제를 미연에 방지한다.

나) 위험 경고: 위험한 날씨 조건에 대한 실시간 경고를 통해 작업 환경을 조정하고, 작업자의 건강을 보호한다.

② 효율적인 자원 관리

가) 작업 일정 조정: 기온과 습도 등 날씨 데이터를 기반으로 작업 일정을 최적화하여 작업자의 건강을 보호하고 작업의 연속성을 유지한다.

나) 비용 절감: 예방 조치를 통해 의료비와 작업 중단 비용을 절감할 수 있다.

다) 정책 및 교육 개선

a. 정책 개발: 예측 결과를 바탕으로 보건 정책과 예방 지침을 개발하여 열사병 예방 효과를 높인다.

b. 교육 자료: 예측 데이터를 활용하여 예방 교육 자료를 개발하고, 근로자의 인식 제고에 도움을 준다.

3) 데이터의 독립 변수와 종속 변수

① 독립 변수

본 예측 모델에서는 다음과 같은 독립 변수를 사용하여 열사병 환자 수를 예측한다.

가) 날짜 및 시간 (Date): 시간적 변화에 따른 계절적 패턴을 반영하기 위해 사용된 변수로, 특정 년, 월, 일, 시간에 따른 기온 및 습도 변화와 열사병 발생의 연관성을 분석한다. 본 프로젝트에서는 날짜 및 시간 데이터를 분할해 년, 월, 일 세 가지 데이터를 독립변수로 사용하였다.

나) 일 최고 기온 (Maximum Temperature): 하루 중 가장 높은 기온을 나타내며, 고온은 열사병 발생의 주요 원인으로 작용한다.

다) 일 평균 기온 (Average Temperature): 일정 기간 동안의 평균 기온으로, 지속적인 고온 상태가 열사병 발생에 미치는 영향을 평가하기 위해 사용된다.

라) 평균 습도 (Average Humidity): 공기의 습도를 나타내며, 높은 습도는 체온 조절을 어렵게 하여 열사병 위험을 증가시킨다.

마) 강수량 합계 (Total Precipitation): 특정 기간 동안의 총 강수량을 기록하며, 비가 오지 않는 날씨가 지속될 경우 열사병 발생 가능성이 증가할 수 있다.

바) 최고-최저 기온차 (Temperature Range): 하루 중 최고 기온과 최저 기온의 차이를 측정하여 기온 변동성을 평가하며, 급격한 기온 변화는 신체 적응을 어렵게 만들어 열사병 위험을 높일 수 있다.

사) 체감온도 (Apparent Temperature): 기온, 습도, 바람 등의 요소를 종합적으로 고려하여 실제로 체감되는 온도를 나타내며, 작업자들이 느끼는 실제 위험 수준을 반영한다.

아) 불쾌지수 (Discomfort Index): 기온과 습도를 바탕으로 신체의 불쾌감을 측정하며, 불쾌지수가 높을수록 열사병 발생 가능성이 높아진다.

자) 전날 열사병 이송 환자 수: 열사병 발생은 연속적인 경향을 보일 수 있어, 전날 환자 수가 많을수록 현재 날에도 높은 발생 확률이 있음을 반영한다.

차) 5 일간 평균 열사병 이송 환자 수: 5 일간 평균 환자 수는 단기적인 기상 변화보다는 장기적인 패턴을 파악하는 데 도움이 되며, 열사병 발생의 전체적인 추세를 고려하는 데 유용하다.

② 종속 변수

가) 열사병 이송 환자 수 (Heatstroke Cases): 종속 변수로, 위의 독립 변수들을 기반으로 예측하려는 주요 변수다. 열사병 이송 환자 수는 하루 동안 발생한 열사병 환자의 총합을 나타내며, 날씨 조건과 직접적인 연관이 있다.

3. 배경지식

1) 데이터 관련 사회 문제 설명

① 산업현장에서의 열사병 위험 증가

최근 산업현장에서 열사병 재해 사례가 빈번하게 발생하고 있다. 특히, 건설, 조선, 농업 등 야외 작업이 많은 분야에서는 고온과 높은 습도 조건에서 열사병 발생 위험이 급격히 증가하고 있다. 열사병은 고온 환경에서 작업하는 근로자들에게 생명을 위협하는 심각한 건강 문제를 초래할 수 있으며, 이는 작업자의 생명과 건강에 직접적인 영향을 미친다.

가) 사례 분석 내용

최근 보도된 사례에 따르면, 2022 년 7 월 대전 신축공사 현장에서 콘크리트 타설 작업을 하던 50 대 노동자 A 씨가 열사병으로 의식을 잃고 쓰러진 후 병원으로 옮겨졌으나 사망한 사건이 발생하였다. 사고 당일 대전 지역의 최고 기온은 33.5 도에 달했으며, 기상청은 '폭염 경보' 특보를 발령할 정도로 극심한 더위가 지속되었다.

이 사건은 작업 환경의 기온과 습도가 높은 상황에서 적절한 예방 조치와 안전 관리가 부족했음을 시사한다. 폭염 경보는 체감온도가 35 도 이상인 상태가 이틀 이상 지속될 것으로 예상될 때 발령된다. A 씨가 수행하던 콘크리트 타설 작업은 지붕이 없는 건물의 꼭대기에서 이루어졌으며, 이는 열사병 발생의 위험을 더욱 증가시킨다.

열사병은 체온 조절 기능을 상실하게 되는 온열 질환 중 가장 위험한 형태로, 고온 환경에서 신경계가 열 자극을 견디지 못해 발생한다. 이와 같은 상황에서는 작업자에게 충분한 휴식시간과 적절한 휴식 공간, 수분 보충이 필요하나, A 씨가 작업을 수행하던 현장에서는 이러한 기본적인 안전 조치가 전혀 지켜지지 않았다. 또한, 열사병 발생 시 신속하게 작업을 중지하고 위험 요인을 제거할 수 있는 매뉴얼이나 대응 계획도 마련되어 있지 않았다.

이에 대해 검찰은 원청업체 대표이사를 중대재해처벌법 위반 혐의로 불구속 기소하였다. 이 사건은 건설현장에서의 열사병 예방과 관련된 안전 관리 시스템의 중요성을 강조하며, 향후 유사 사건을 방지하기 위한 강력한 안전 관리와 정책 개선이 필요함을 시사한다.

2) 데이터 기반 접근의 필요성

열사병 이송 환자 수 예측 프로그램을 효과적으로 개발하기 위해서는 다양한 데이터에 대한 접근이 필수적이다. 정확한 예측을 위해 기온, 습도, 강수량 등 기상 데이터뿐 만 아니라, 환자의 이송 기록과 같은 건강 관련 데이터도 필요하다. 이러한 데이터들은 열사병 발생의 패턴을 파악하는 데 중요한 역할을 하며, 모델의 예측 정확도를 높이는 데 기여한다. 데이터 접근을 통해 실시간 및 축적된 정보를 분석함으로써, 열사병의 위험을 사전에 예측하고 적절한 대응 조치를 취할 수 있는 기반을 마련할 수 있다. 따라서 데이터의 수집과 활용은 열사병 예측 프로그램의 신뢰성과 효과성을 보장하는 핵심 요소이다.

4. 개발 내용

본 연구에서 사용된 데이터는 기상 및 환경 조건과 열사병 이송 환자 수와 관련된 정보로 구성되어 있다. 훈련 데이터(train_df)는 총 41 개의 열을 가지고 있으며, 테스트 데이터(test_df)는 39 개의 열로 구성되어 있다. 테스트 데이터는 21 열인 월(Month)과 41 열인 연도(Year) 열이 누락되어 있어 해당 열을 분석에서 제외하였다.

1) 데이터 간 상관관계 설명

분석을 통해 열사병 이송 환자 수와 유의미한 상관관계를 보이는 변수들을 선별하였다. 선정된 독립변수는 '년(Year)', '월(Month)', '일(Day)', '일 최고 기온', '일 평균 기온', '평균 습도', '강수량 합계', '최고-최저 기온차', '체감온도', '불쾌지수', '전일의 이송 인원수', '이송 인원수 이동 평균(5 일간)'으로, 총 12 개의 변수이다. 상관관계는 (그림 1)과 같다. 산점도로 값을 표현하고, 회귀선을 그려 한 눈에 알아보기 쉽도록 나타냈다.

① 년(Year): 개별 상관관계 분석에서는 연도와 종속변수 간에 큰 상관관계가 없는 것으로 보인다. 하지만 연도 데이터를 포함하여 학습을 진행하였을 때 정확도가 유의미하게 증가하는 결과가 나타났다.

② 월(Month): 일년 중 가장 더운 시기인 7 월과 8 월에 열사병 환자 수가 가장 많이 발생한다는 상관관계를 보인다.

③ 일(Day): 연도와 마찬가지로 포함하여 학습을 진행하였을 때 정확도가 유의미하게 증가하는 결과가 나타났다.

④ 일 최고 기온: 일 최고 기온이 높아질수록 열사병 환자 수가 증가한다. 또한 일 최고기온이 30 도를 넘기면 열사병 환자 수가 기하급수적으로 증가한다는 것을 알 수 있다.

⑤ 일 평균 기온: 일 평균 기온이 높아질수록 열사병 환자 수가 증가한다. 또한 일 평균 기온이 25 도를 넘기면 열사병 환자 수가 기하급수적으로 증가한다는 것을 알 수 있다.

⑥ 평균 습도: 평균 습도가 70% 근처일 때 열사병 환자 수가 가장 많다. 습도가 80% 이상일 땐 열사병 환자 수가 크게 줄어드는데, 해당 습도에서는 비가 오기 때문인 것으로 파악하였다. 따라서 습도는 높지만 비가 오지 않는 수준에서 가장 열사병 환자가 많이 발생할 것으로 보인다.

⑦ 강수량 합계: 대부분의 열사병 환자는 강수량이 0 일 때 발생한다.

⑧ 최고-최저 기온차: 일교차에 해당하는 독립변수로, 기온차가 6 도 이상 10 도 미만일 때 열사병 환자 수가 많다. 이 일교차는 7 월과 8 월의 일교차와 일치한다.

⑨ 체감온도: 체감온도가 높아질수록 열사병 환자 수가 증가한다. 또한 체감온도가 25 도를 넘기면 열사병 환자 수가 기하급수적으로 증가한다는 것을 알 수 있다.

⑩ 불쾌지수: 불쾌지수가 높아질수록 열사병 환자 수가 증가한다. 또한 불쾌지수가 75 를 넘기면 열사병 환자 수가 기하급수적으로 증가한다는 것을 알 수 있다.

⑪ 전날의 열사병 환자 이송 인원 수: 양의 상관관계이다. 이는 다음날 발생할 열사병 환자 수를 예측하는 데 전날 발생한 열사병 환자의 수가 중요한 역할을 한다는 것을 알 수 있다.

⑫ 이송 인원수 이동 평균(5 일간): 역시 양의 상관관계이다.

위와 같은 상관관계 분석 결과를 산점도 및 회귀선 그래프로 나타냈다. 또한 밀도그래프(그림 2)를 추가하여 한 눈에 보기 쉽도록 시각화 작업을 진행하였다.

2) 예측 목표 및 변수 설정

본 연구의 목표는 주어진 기상 및 환경 데이터를 기반으로 특정 일의 열사병 이송 환자 수를 예측하는 것이다. 독립변수는 앞서 상관관계 분석을 통해 선정한 12 가지이다. 독립변수의 선정기준은 다음과 같다.

① 종속변수와 상관관계가 있을 것

② 일기예보, 기상청 홈페이지 등에서 쉽게 얻을 수 있는 날씨 정보일 것

③ 날씨 정보 외에 예측 모델의 완성도를 높일 수 있는 유의미한 변수일 것

선정된 모든 독립변수는 항목 ①에 해당하며, 항목 ②에 해당하는 10 가지와 항목 ③에 해당하는 2 가지 독립변수로 이루어져 있다.

종속변수는 해당 독립변수들로 결과를 예측할 '열사병 이송 환자 수' 이다.

3) Gradient Boosting 모델 선정 이유

열사병 발생과 관련된 변수들(예: 기온, 습도, 불쾌지수 등)은 비선형 관계를 가질 수 있다. 예를 들어, 기온과 습도가 높아질수록 열사병 발생 확률이 기하급수적으로 증가할 수 있다. 이러한 비선형성을 효과적으로 모델링하기 위해 Gradient Boosting 모델을 선정하였다.

Gradient Boosting 모델은 순차적으로 약한 학습기를 학습시켜 오류를 줄여나가는 방법으로, 성능 향상에 효과적이다. 이 모델은 비선형 관계를 잘 모델링할 수 있으며, 높은 정확도를 제공하는 동시에 과적합을 방지하는 기능이 있다.

4) 성능 비교를 위한 Random Forest 모델 선정 이유

Random Forest 는 여러 개의 결정 트리를 기반으로 하는 앙상블 방법으로, 각 트리가 예측을 독립적으로 수행하고 최종 예측을 평균화하여 성능을 향상시킨다. 이 모델은 변수 간의 복잡한 상호작용을 잘 포착하며, 데이터의 잡음에 강하고 과적합을 방지하는 특성이 있다. 그러나 단점으로는 많은 트리를 생성하기 때문에 계산 비용이 높아질 수 있으며, 변수 중요도를 해석하기 어렵다는 점이 있다. 또한, 매우 큰 데이터셋에서는 학습 속도가 느려질 수 있다는 한계가 있다. 성능 비교 모델로서는 적합하지만, 최종 모델로는 적합하지 않다고 판단하였다.

5) 사용할 성능 지표 및 성능 지표 선정 이유

① MSE (Mean Squared Error): MSE 는 예측 값과 실제 값 간의 차이를 제곱하여 평균을 낸 것이다. 값이 작을수록 모델의 예측 성능이 좋다고 평가할 수 있다. 이 지표는 예측 오차의 제곱에 비례하여 큰 오차에 대해 더 큰 페널티를 부여하기 때문에, 모델이 큰 오차를 줄이는 데 유리하다.

② RMSE (Root Mean Squared Error): RMSE 는 MSE 의 제곱근을 취한 것이다. MSE 에 비해 더 직관적으로 이해할 수 있는 지표로, 오차의 단위와 예측 값의 단위가 동일하여 해석이 용이하다. 값이 작을수록 예측이 실제 값에 가까운 것을 의미한다.

③ MAE (Mean Absolute Error): MAE 는 예측 값과 실제 값 간의 차이의 절대값의 평균이다. 값이 작을수록 모델의 예측 성능이 좋다고 볼 수 있다. MAE 는 오차의

크기에 대해 균등하게 가중치를 부여하여, 전체 오차의 평균적인 크기를 직접적으로 반영한다.

④ R^2 (R-squared): R^2 는 모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표이다. 1 에 가까울수록 모델이 데이터의 변동성을 잘 설명하는 것으로 평가된다. 이 지표는 모델의 설명력을 직관적으로 제공하며, 예측 성능을 평가하는 데 유용하다.

5. 개발 결과

1) Gradient Boosting 모델

① Gradient Boosting 모델의 KFold 교차검증 결과(그림 3)

가) MSE (약 177.74): 교차 검증에서 평균적인 오차 제곱이 약 177.12 임을 나타낸다. MSE 가 낮을수록 예측이 실제 값과 가깝다.

나) RMSE (약 13.22): MSE의 제곱근으로, 예측 오차의 평균 크기를 나타낸다. 약 13.20은 모델의 평균적인 오차를 나타낸다.

다) MAE (약 7.42): 예측과 실제 값 간의 절대적 평균 오차가 약 7.42이다. 이는 예측이 실제 값과 평균적으로 약 7.42만큼 차이가 난다는 의미이다.

라) R^2 (0.89): 모델이 데이터의 약 89%를 설명할 수 있다는 뜻이다. 1에 가까울수록 모델의 설명력이 좋다.

② Gradient Boosting 모델의 성능 평가 결과(그림 3)

가) MSE (1405.77): 교차 검증에서 평균적인 오차 제곱이 약 1405.77임을 나타낸다. MSE가 낮을수록 예측이 실제 값과 가깝다.

나) RMSE (37.49): MSE의 제곱근으로, 예측 오차의 평균 크기를 나타낸다. 약 37.49은 모델의 평균적인 오차를 나타낸다.

다) MAE (16.62): 예측과 실제 값 간의 절대적 평균 오차가 약 16.62이다. 이는 예측이 실제 값과 평균적으로 약 16.62만큼 차이가 난다는 의미이다.

라) R^2 (0.81): 모델이 데이터의 약 81%를 설명할 수 있다는 뜻이다. 1에 가까울수록 모델의 설명력이 좋다.

2) Random Forest 모델과 성능 비교

① Random Forest 모델의 KFold 교차검증 결과(그림 4)

가) MSE (186.28): 비교적 낮은 값으로, 교차검증 데이터에서 모델이 평균적으로 적당히 좋은 성능을 보인다.

나) RMSE (13.48): 예측값과 실제값 간의 오차가 약 13.48 단위이며, 낮은 편이지만 Gradient Boosting 모델보다 약간 높게 나타난다.

다) MAE (7.47): 극단적인 오차에 영향을 덜 받았을 때의 오차이다.

라) R^2 (0.88): 값이 높은 수치로 나타나며, Random Forest 모델이 열사병 환자 예측을 위한 데이터와 잘 맞아 떨어지는 모델임을 알 수 있다.

② Random Forest 모델의 성능 평가 결과(그림 4)

가) MSE (1481.24): Random Forest 모델의 MSE 가 1481.24 으로 나타났다. 이는 테스트 데이터에서 모델이 예측하는 데 있어서 오차가 크다는 것을 의미한다. MSE 값이 높다는 것은 모델의 예측 오차가 비교적 크며, 성능이 다소 부족함을 시사한다.

나) RMSE (38.49): 테스트 데이터에서 RMSE 가 38.49 로 나타났다. 이는 예측 오차의 평균적인 크기가 38.49 로 비교적 높다는 것을 의미한다. RMSE 값이 크다는 것은 모델의 예측이 실제 값과 상당한 차이를 보임을 나타낸다.

다) MAE (17.09): 테스트 데이터에서 MAE 가 17.09 으로 나타났다. 이는 예측 값과 실제 값 간의 평균적인 절대 오차가 17.09 으로 증가했음을 의미한다. MAE 값이 높다는 것은 모델이 예측에서 절대적인 오차를 줄이는 데 한계가 있음을 시사한다.

라) R^2 (0.80): Random Forest 모델의 R^2 값이 0.79 로 나타났다. 이는 모델이 테스트 데이터의 약 80%를 설명할 수 있음을 의미한다. KFold 교차 검증 결과보다 낮지만 여전히 좋은 설명력을 보이며, 데이터의 변동성을 상당히 잘 설명하고 있다.

3) 머신러닝 모델의 성능 결과에 대한 해석

이러한 성능 지표를 통해 Gradient Boosting 과 Random Forest 모델 모두 테스트 데이터에서 안정적이고 신뢰할 수 있는 예측 성능을 보이고 있음을 알 수 있다. 결과를 시각화한 그래프를 보면(그림 5)(그림 6), 특히 Gradient Boosting 모델이 다소 더 나은 성능을 보인다. 시각화는 총 4 가지의 그래프를 이용하여 나타냈다.

① 산점도: 실제 값과 예측 값 간의 관계를 시각화하여 모델의 성능을 평가한다. 대각선에 점들이 가까이 모여 있을수록 예측이 실제 값과 가까운 것이며, 대각선에서 멀리 떨어진 점들은 예측이 잘못된 경우이다.

② 실제 값-예측 값 분포 히스토그램: 실제 값과 예측 값의 분포를 비교하여 예측 값이 실제 값을 얼마나 잘 반영하는지 평가한다. 두 히스토그램이 서로 가까이 모여 있을수록 예측 값의 분포가 실제 값의 분포와 잘 맞는 것이다.

③ 선 그래프: 예측 값의 변화를 점선 그래프와 마커로 표시한다. 두 선이 비슷한 패턴을 보일수록 예측이 잘 이루어진 것이다.

④ 오차 분포 히스토그램: 예측 오차의 분포를 시각화하여 모델의 예측 오류를 분석한다. 오차의 분포가 중심에 모여 있을수록 예측이 전반적으로 잘 이루어진 것이다.

Gradient Boosting 학습 모델과 Random Forest 학습 모델의 그래프를 비교하였을 때, Gradient Boosting 학습 모델이 비교적 실제 값과 유사하게 예측한다는 것을 알 수 있다. 특히 세 번째 선 그래프에서 예측 값의 패턴이 실제 값의 패턴과 유사하게 나타나는 것을 볼 수 있다. 실제 열사병 이송 환자 수가 크게 높아진 지점에서 예측 값도 동일하게 높아지며, 반대로 환자 수가 크게 낮아진 지점에서도 예측 값 역시 동일한 패턴을 보인다. 또한 네 번째 오차 분포 히스토그램에서 오차 분포가 중심에 모여 있어 예측이 잘 이루어져 있음을 알 수 있다.

이는 Gradient Boosting 이 데이터의 복잡한 패턴을 더 효과적으로 학습했음을 나타내며, 최종적으로 이 모델을 선택하게 된 이유이다. 두 모델 모두 긍정적인 결과를 도출했지만, 정확성과 예측력에서 우수한 Gradient Boosting 모델이 실제 데이터 예측에 있어 더 유용한 도구임을 시사한다.

6. 결론

1) 머신러닝 모델 개발에 관한 간략한 요약 및 결과 설명

본 프로젝트는 건설현장에서 열사병 발생을 예측하기 위해 Gradient Boosting 모델을 개발하였다. 주요 변수로는 기온, 습도, 불쾌지수 등을 포함하였고, K Fold 교차 검증을 통해 모델의 성능을 평가하였다.

① 데이터 처리 및 모델링

가) 데이터 처리: 기온, 습도 등 날씨 데이터와 열사병 이송 환자 수를 포함한 데이터를 전처리하고, 주요 변수를 선정하였다.

나) 모델링: Gradient Boosting 과 Random Forest 를 적용하여 성능 비교 후 정확도가 더 높은 모델을 선정하여 예측 모델을 구축하였다.

② 성능 평가 결과

Gradient Boosting: MSE 1405.77, RMSE 37.49, MAE 16.62, R^2 0.81

③ 결론: 모델의 성능 지표는 전체적으로 우수하며, 특히 Gradient Boosting 모델은 낮은 MAE 와 RMSE 값으로 열사병 환자 수 예측에서 매우 신뢰할 수 있는 결과를 제공하였다. 이러한 결과는 건설현장에서의 작업 안전성을 강화하고, 열사병 예방을 위한 실질적인 조치를 취하는 데 기여할 수 있을 것으로 기대된다.

④ 향후 개선 방향: 모델의 예측 성능을 더욱 향상시키기 위해, 데이터의 계절적 변동성과 기후 변화에 대한 적응력을 높이고, 최신 기후 데이터를 반영하여 모델을 지속적으로 업데이트할 필요가 있다. 또한, 추가적인 데이터 수집과 피드백을 통해 모델의 신뢰도를 더욱 높일 수 있을 것으로 보인다.

2) 개발 의의

본 모델 개발은 건설현장에서 열사병 발생을 예측하고 예방하여 작업자의 안전과 건강을 보호하는 데 중요한 가치를 제공한다. 예측 결과를 통해 위험이 높은 작업 환경을 미리 경고하고, 관리자와 작업자가 적절한 예방 조치를 취할 수 있도록 지원함으로써, 사고로 인한 작업 중단과 인명 피해를 줄일 수 있다.

또한, 본 모델은 데이터 기반으로 열사병 발생 가능성을 과학적으로 분석하여 안전 관리의 수준을 높인다. 경험이나 직관에 의존하던 기존 방법을 대체하고, 보다 정교하고 객관적인 판단을 가능하게 한다. 이는 작업 효율성 증대와 건설업계의 안전 수준 향상에 기여할 것이다.

이 모델은 건설현장 외에도 농업, 물류, 야외 행사 등 다양한 산업에서 폭염으로 인한 건강 문제를 해결하는 데 활용될 수 있다. 따라서, 본 프로젝트는 폭염 대응 및 열사병 예방에 대한 새로운 표준을 제시하는 데 의미가 있다.

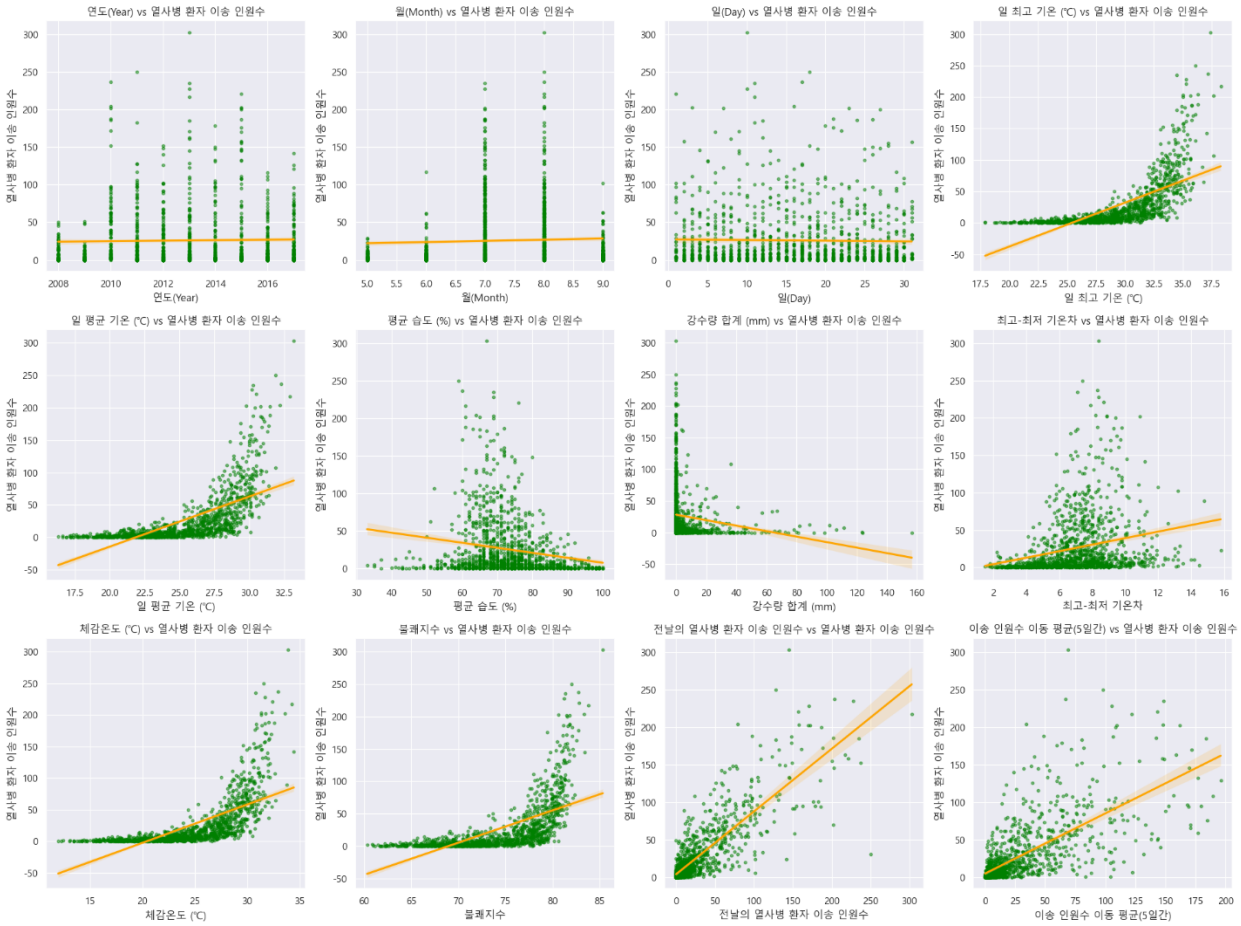
3) 머신러닝 모델의 한계

기후 변화와 갑작스러운 날씨 변동은 예측 변수의 변동성을 증가시킬 수 있다. 예를 들어, 예측 시점에 갑작스러운 기온 상승이나 강수량의 급격한 변화는 모델이 학습한 패턴과 다를 수 있으며, 이로 인해 모델의 예측 정확도에 영향을 줄 수 있다. 기후 변화로 인해 장기적인 기온 상승이 지속되면, 기존 모델이 예측하는 패턴과 실제 패턴 간의 불일치가 발생할 수 있다.

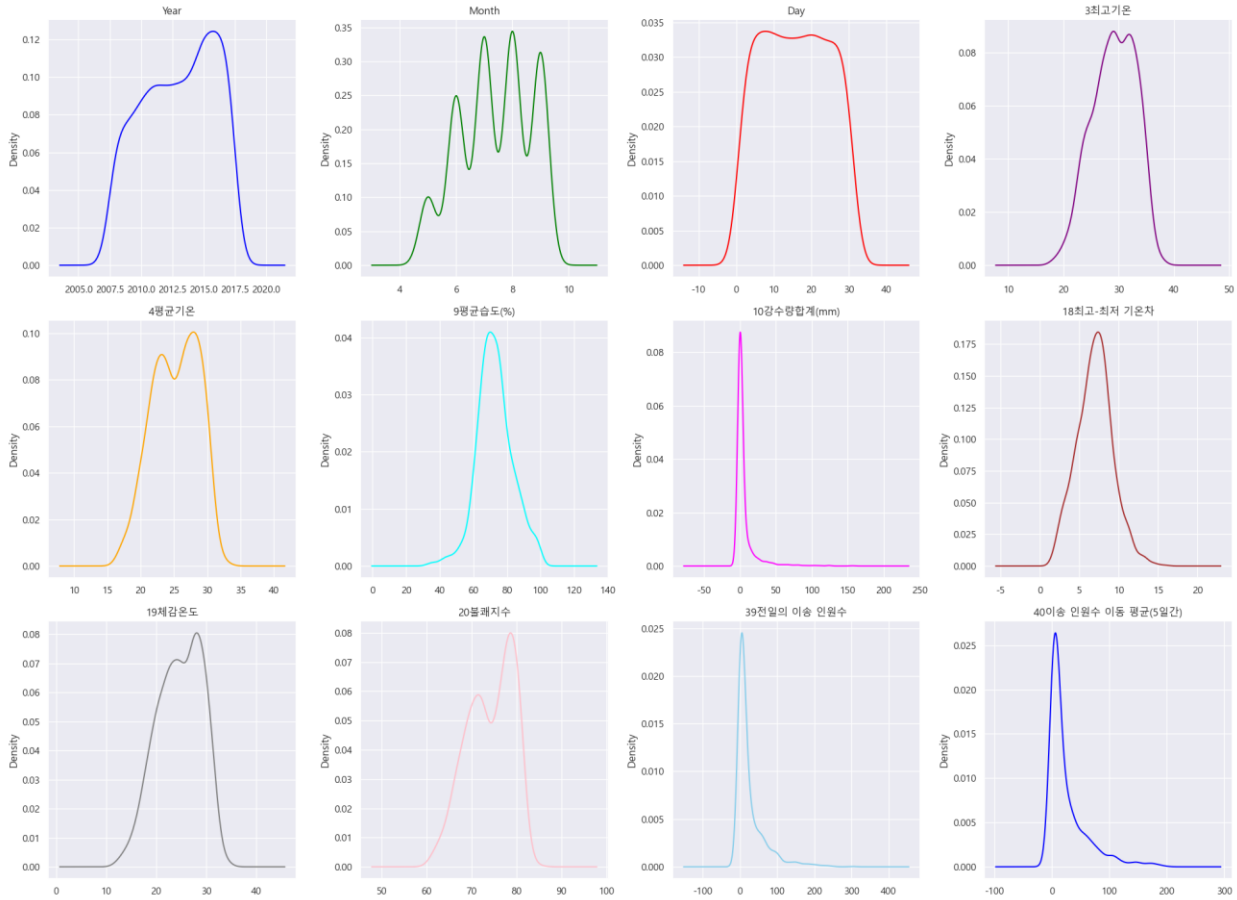
이를 해결하기 위해 기후 변화와 날씨 변동성을 고려하여, 최신 기후 데이터를 반영하고 장기적인 기후 트렌드를 모델에 통합하는 것이 필요하다.

예를 들어, 예측 모델에 최신 기후 데이터와 변동성 요소를 추가하고, 기후 변화에 대한 적응력을 높이기 위해 주기적인 모델 업데이트를 수행하여 불확실성을 줄일 수 있을 것으로 보인다.

<그림 및 그래프>



(그림 1) 각 독립변수와 종속변수 간의 상관관계를 나타낸 산점도 및 회귀선



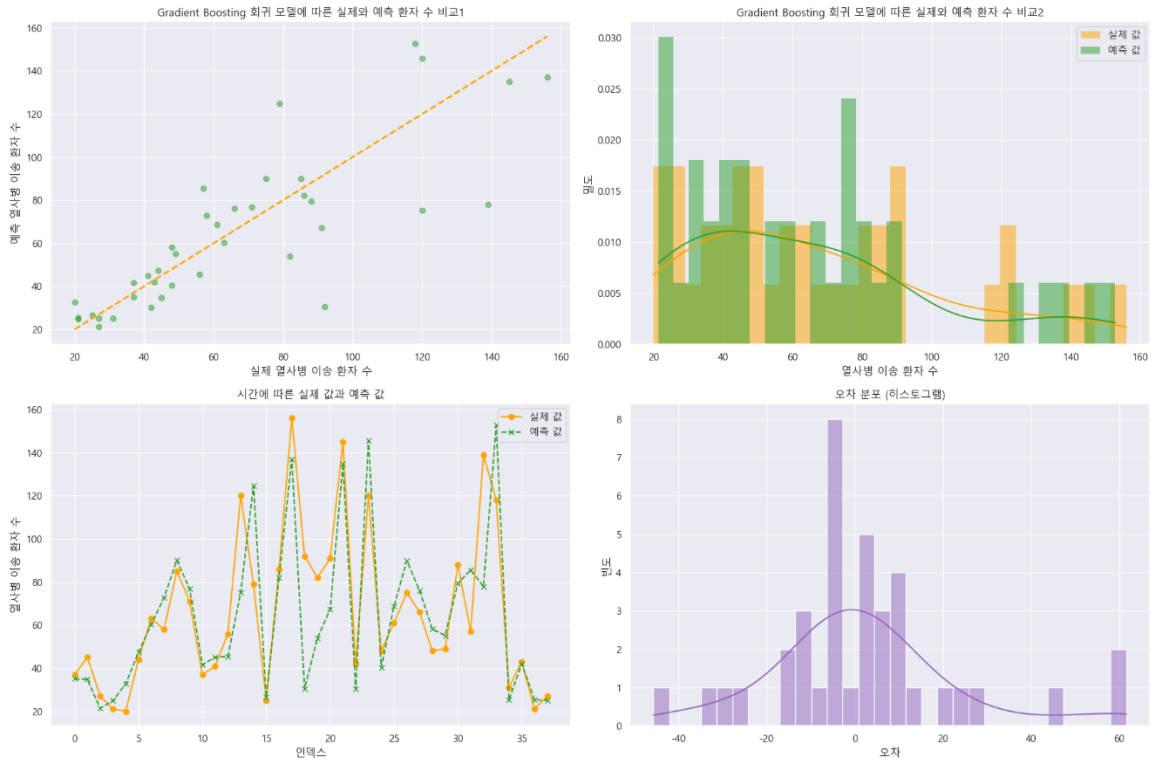
(그림 2) 각 독립변수와 종속변수 간의 상관관계를 나타낸 밀도 그래프

```
-----Gradient Boosting KFold 교차검증 결과-----
KFold Cross-Validation MSE: 177.74
KFold Cross-Validation RMSE: 13.22
KFold Cross-Validation MAE: 7.42
KFold Cross-Validation R^2: 0.89


-----Gradient Boosting 성능평가 결과-----
MSE: 1405.77
RMSE: 37.49
MAE: 16.62
R^2: 0.81

Process finished with exit code 0
```

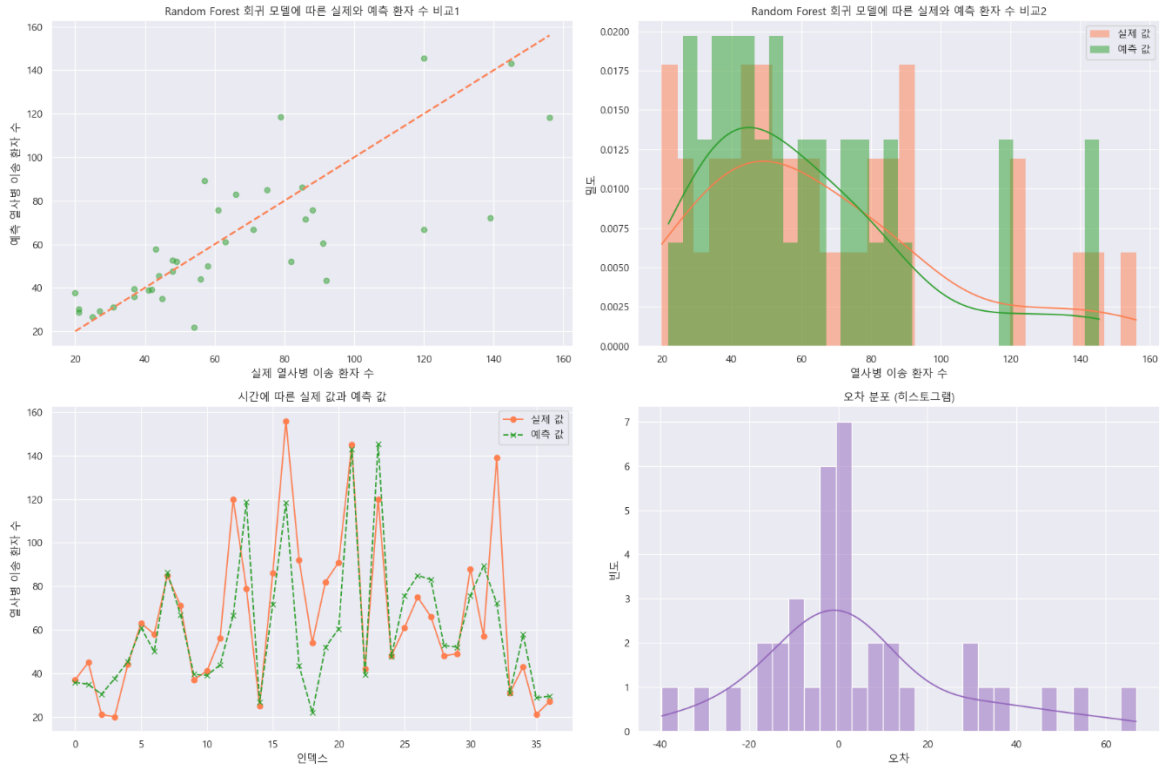
(그림 3)Gradient Boosting 학습 모델

```
-----Random Forest KFold 교차검증 결과-----
KFold Cross-Validation MSE: 186.28
KFold Cross-Validation RMSE: 13.48
KFold Cross-Validation MAE: 7.47
KFold Cross-Validation R^2: 0.88


-----Random Forest 성능평가 결과-----
MSE: 1481.24
RMSE: 38.49
MAE: 17.09
R^2: 0.80

Process finished with exit code 0
```

(그림 4)Random Forest 학습 모델



(그림 5) Gradient Boosting 학습 모델



(그림 6) Random Forest 학습 모델