JOURNAL OF RADIATION INDUSTRY

Technical article

17(4) : 551-557 (2023)
https://doi.org/10.23042/radin.2023.17.4.551

방사선량률 예측을 위한 기계학습 기반 모델 개발 및 최적화 연구

이시현[1], 이홍연[1,*], 염정민[1]

[1](주)알엠택

Machine Learning Based Model Development and Optimization for Predicting Radiation

SiHyun Lee[1], HongYeon Lee[1,*] and JungMin Yeom[1]

[1]RMTEC Co., Ltd., 25, Hanamsandancheonbyeonjwa-ro, Gwangsan-gu, Gwangju-si 62215, Republic of Korea

Abstract In recent years, radiation has become a socially important issue, increasing the need for accurate prediction of radiation levels. In this study, machine learning-based models such as Multiple Linear Regression (MLR), Random Forest (RF), XGBoost, and LightGBM, which predict the dose rate by time (nSv h^{-1}) by selecting only important variables, were used, and the correlation between temperature, humidity, cumulative precipitation, wind direction, wind speed, local air pressure, sea pressure, solar radiation, and radiation dose rate (nSv h^{-1}) was analyzed by collecting weather data and radiation dose rate for about 6 months in Jangseong, Jeollanam-do. As a result of the evaluation based on the RMSE (Root Mean Squared Error) and R-Squared (R-Squared coefficient of determination) scores, the RMSE of the XGBoost model was 22.92 and the R-Squared was 0.73, showing the best performance among the models used. As a result of optimizing hyperparameters of all models using the GridSearch method and comparing them by adding variables inside the measuring instrument, it was confirmed that the performance improved to 2.39 for RMSE and 0.99 for R-Squared in both XGBoost and LightGBM.

Key words: Machine learning, Dose rate, Optimization, RMSE

1. 서 론

방사능은 인체의 건강과 환경에 대한 중요한 이슈로 최근 후쿠시마 원전 방사성 오염수 방류 문제로 인한 국내외에서의 방사선/능 관련 우려와 논의 등 사회적 및 문화적 문제가 증가하고 있다. 이로 인해 국가 및 지자체 단위에서 방사선/능 관리와 안전에 대한 논의와 대책 마련이 활발하게 진행되고 있다. 특히 방사선/능 노출 및 방사능 농도 예측은 방사선/능 안전 및 관련 정책 수립에 있어 핵심적인 요소지만 방사선/능은 다양한 현상과 환경 조건에 따라 예측하기 어렵다. 이러한 이유로 방사선/능 예측 모델의 중요성이 점차 높아지고 있어 이와 관련한 연구가 활발히 이루어져야 한다. 그러나 현재 사용 중인 방사선/능 예측 모델은 예측의 정확성에 한계가 있어, 보다 정확하고 효율적인 방사선/능 예측 모델의 개발이 필요한 상황이다. 그러므로 본 연구에서는 방사선/능 예측 모델의 개선을 목표로 전라남도 장성 지역에서 2022년 9월부터 2023년 3월까지 수집한 방사선량률 데이터와 기상청에서 제공한 데이터를 활용하였으며, 기온, 습도, 누적 강수량, 풍향, 풍속, 현지 기압, 해면 기압 등과 선량

http://www.ksri.kr/

*Corresponding author. HongYeon Lee
Tel. +82-62-236-6740 Fax. +82-62-236-6741 E-mail. tlgus7668@gmail.com
Received 7 December 2023 Revised 18 December 2023 Accepted 19 December 2023

률(nSv h^{-1})을 변수로 선택하여 사용하였다. 또한 기계학습 알고리즘인 Random Forest, LightGBM (Light Gradient Boosting Model), XGBoost (eXtreme Gradient Boosting), 다중선형회귀를 활용하여 방사선량률(nSv h^{-1})을 예측하였다. 그리고 각 알고리즘의 성능을 비교 및 평가를 통해 방사선/능 예측에 있어 보다 효과적인 모델이 무엇인지, 각각의 모델에 최적화 방법을 사용하여 하이퍼파라미터(Hyperparmeter)를 선정하여 모델의 성능 향상, 측정 내부 요인이 선량값에 미치는 영향 등을 평가하였다. 이를 통해 보다 정확하고 효율적인 방사능 예측을 위한 개선사항과 새로운 방법론을 제시하였다.

2. 재료 및 방법

2.1. 기상 데이터

기상청은 각 관측소에서 분 단위로 측정된 기상 정보를 제공하며, 이 정보에는 관측소의 위치와 측정 시간을 기반으로 한 온도, 습도, 풍향, 풍속, 기압 등의 기상 데이터가 포함되어 있다. 따라서 본 연구에서는 방사선/능 예측을 위한 기상자료로 기상자료개방포털에서 CSV (Comma Separated Values; 데이터 직렬화 포맷 중 표 형태의 데이터를 저장하는 파일 형식의 확장자명) 형식으로 제공하는 전라남도 장성의 종관기상관측(Automated Synoptic Ovserving System, ASOS) 2022년 9월 1일부터 2023년 3월 14일까지(약 6개월) 데이터를 활용하였다[1].

2.2. 선량 데이터

또한 방사선 측정기를 통해 전라남도 장성 지역의 대기 중 방사선량률를 측정한 데이터를 사용하였다[2]. 측정된 데이터에는 방사선량률(nSv h^{-1}), 장비온도(°C), 전광판 온도(°C), 습도, 배터리 잔량, 일시 등의 정보를 포함하여 5분 단위로 측정된 데이터셋을 제공받아 활용하였다.

본 연구에서는 기상 데이터 수집기간과 같은 기간의 데이터를 활용하였으며, 일부 데이터의 경우 데이터 입력값이 기록되지 않았거나 관측되지 않은 결측값에 대해서는 데이터 품질차원에서 신뢰성이 없다고 판단하여 제외하였다.

2.3. 데이터 전처리

방사선량률 데이터와 기상청의 기상 데이터의 형식이 상호 다르기 때문에, 두 데이터 형식을 통일화하는 단계를 거친 후, 데이터 정규화를 수행하였다. 사용한 데이터 중 누적 강수량(mm)의 경우 데이터 입력값이 기록되지 않았거나 관측되지 않은 결측값(NaN, Not a Number)이 있었다. 따라서 이러한 결측값 대신 0으로 처리하여 데이터 품질을 고려하였으며, 이외에 나머지 변수의 결측값은 결측치가 발생한 시간 이전 시간 값으로 대체하여 사용하였다. 그리고 '기온(°C)', '누적강수량(mm)', '현지기압(hPa)', '해면기압(hPa)', '선량률(nSv h^{-1})', '풍향(deg)', '풍속(m/s)' 등 모든 변수를 정확한 예측을 위해 수식(1)의 Min-Max Scale 방식을 사용하여 0과 1 사이의 값으로 정규화한 후 선량 데이터와 기상 데이터를 병합하였다.

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

이와 관련하여 Table 1은 본 연구에서 활용한 데이터를 종합하여 나타내었다. 또한 Table 1과 같이 본 연구에서 활용한 데이터에 측정기 내부 요인(장비온도, 전광판온도, 습도)이 선량값에 미치는 영향을 평가하기 위해 해당 변수를 추가하여 Table 2에 나타내었다[3].

Table 1. Used data

Classification	Feature	Measurement period
Radiation data	nSv h^{-1}	5 minute (6 months)
Weather data	Temperatrue (°C) Accumulated precipitation (mm) Wind direction (deg) Wind speed (m s^{-1}) Local pressure (hPa) Sea level pressure (hPa) Humidity (%)	

2.4. 모델

2.4.1. 예측 모델

실험을 목적으로 수치 예측에 주로 사용되는 머신러닝 알고리즘은 MLR, RF, XGBoost, LightGBM 등이 있다.

2.4.1.1. 다중선형회귀

다중선형회귀 방법은 종속 변수와 두 개 이상의 독립

변수 간의 선형 관계를 모델링하고 예측하는데 사용된다. 이 모델은 여러 독립 변수를 고려하여 종속 변수를 예측하며, 이러한 관계를 선형 함수로 표현하고, 모델 훈련 과정에서 독립 변수의 계수를 추정함으로써 각 독립 변수가 종속 변수에 미치는 영향을 정량화하며, 이를 통해 새로운 데이터 포인트에 대한 예측을 수행할 수 있는 방법이다.

2.4.1.2. Random Forest

Random Forest 모델은 고차원 데이터나 다수의 입력 변수를 다룰 때 정확한 예측을 수행하면서도 과적합(Over-fitting) 문제를 효과적으로 해결하기 위한 방법이다[4]. 이 모델은 앙상블(ensemble) 학습 기반으로, 여러 개의 분류기를 결합하여 더 정확한 예측을 제공하고, 트리(tree)를 생성하는 과정에서 일부 훈련 데이터셋(Training dataset)을 무작위로 선택하여 배깅(bagging)을 수행한다. 이렇게 생성된 여러 개의 트리에서 나온 결과 중 가장 빈도가 높은 값을 최종 예측 결과로 선택하는 방법이다.

2.4.1.3. XGBoost

XGBoost는 다수의 결정 트리를 결합하여 예측을 수행하는 알고리즘으로, 과적합 문제를 완화하는 방법으로 병렬 연산을 활용하여 데이터 처리 속도를 향상시키며, 결정 트리의 복잡성을 제한하는 패널티 항을 추가하여 과적합을 방지한다. 또한, Loss Function을 최소화하여 모델을 학습하고 최적화하는 방법이다[5].

2.4.1.4. LightGBM

LightGBM은 Gradient Boosting 알고리즘 중 하나로, 경량이면서 빠른 학습 및 예측 속도를 가지며 대용량 및 고차원 데이터셋에서 높은 성능을 제공한다. 이 모델은 leaf 중심 트리 분할 방식을 사용하여 효율적으로 데이터를 처리하고, 범주형 특징을 자동으로 다룰 수 있어 전처리 과정을 단순화한다. XGBoost와 같은 GBDT (Gradient Boosting Decision Tree) 기법을 기반으로 속도를 보완하였고, GBDT의 각 노드에서 분기점이 나뉠 때, 잘 맞는 노드를 기준으로 분리하며, 잘 맞지 않는 노드는 분기점으로 선택하지 않는 방법이다[6].

2.4.2. 예측 방법

본 논문에서는 방사선량률(nSv h^{-1}) 예측을 위해 Table 1의 변수값과 Table 2의 측정기 내부 요인 데이터를 추가하여 2022년 9월 1일부터 2023년 3월 14일까지 5분 단위 데이터 총 47,076개를 사용하여 각 모델별로 실험을 진행하여 성능을 비교하였다.

Table 2. Data with added internal factors of the measuring instrument

Classification	Feature	Measurement period
Radiation data	nSv h^{-1}	5 minute (6 months)
Weather data	Temperatrue (°C)	
	Accumulated precipitation (mm)	
	Wind direction (deg)	
	Wind speed (m s^{-1})	
	Local pressure (hPa)	
	Sea level pressure (hPa)	
	Humidity (%)	
Measurement data	Measurement temperature (°C)	
	Measurement humidity (%)	

예측 모델은 5분 단위로 수집한 선량 및 기상 데이터를 7 : 3의 비율로 분할하여 32,953개의 훈련 데이터(Training Data)를 사용하여 모델을 학습하였으며, 이 모델의 성능 평가에는 14,123개의 검증 데이터(Validation Data)를 활용하였다.

Fig. 1에 본 연구에서 진행한 예측 방법을 시각적으로 나타내었고, 측정기 내부 요인 변수를 포함하여 모델의 성능에 미치는 영향을 추가적으로 분석하였다. 이와 관련하여 Fig. 2에서 측정기 내부 요인을 추가했을 때의 과정을 나타내었다.

예측 성능을 평가하는 지표로는 회귀 모델에서 일반적으로 사용되는 RMSE (Root Mean Squared Error, 평균제곱근오차)와 R-Squared (결정계수)를 선택하였다.

$$RMSE = \sqrt{\frac{\sum (y - \hat{y})^2}{n}} \tag{2}$$

수식 (2) RMSE는 예측 값과 실제 값 사이의 차이를 제곱하여 전체 데이터 개수로 나눈 후 루트를 씌운 값으로, 모델의 예측 정확도를 나타내는 지표이다. R-Squared는

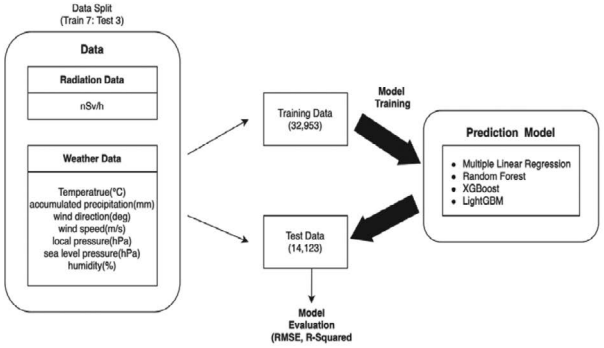

Fig. 1. How to predict data.

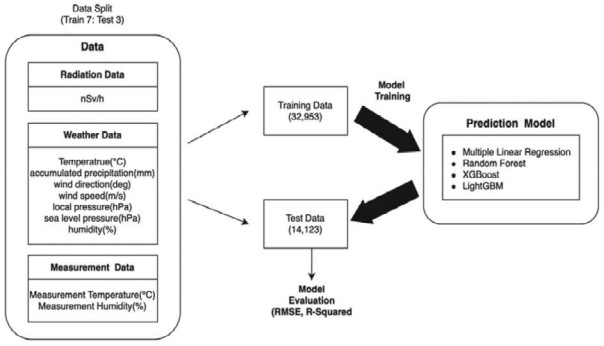

Fig. 2. Data prediction method that adds internal factors of the measuring instrument.

모델이 종속 변수의 분산을 얼마나 잘 설명하는지를 나타내며, 높을수록 모델의 설명력이 좋다는 것을 의미한다.

이와 같은 방법을 모든 알고리즘의 예측 성능 평가 지표로 활용하였다.

2.4.3. 예측 모델 최적화

Random Forest, LightGBM, XGBoost의 경우, 각 모델의 성능은 하이퍼파라미터 (Hyperparmeter)에 따라 다를 수 있다. 그러므로 하이퍼파라미터는 모델의 학습 및 예측 과정을 제어하는 매개변수로, 이러한 하이퍼파라미터를 최적화하여 모델의 성능을 향상시킬 수 있다. 따라서 본 연구에서는 각 모델의 최적 하이퍼파라미터를 선정하기 위해 가능한 하이퍼파라미터 조합을 사전에 정의한 범위에서 모두 시도하여 파라미터 결과가 개선되는 경우에만 이전 파라미터 값을 갱신하여 최적의 파라미터를 선정하는 방법인 Grid Search 기법을 사용하여 RF, XGBoost, LightGBM 파라미터를 최적화하였다.

Table 3. Random forest hyperparameter-1

Parameter	Optimization value
max_depth	6
n_estimators	100
min_samples_leaf	2
min_samples_split	2

Table 4. Random forest hyperparameter-2

Parameter	Optimization value
max_depth	6
n_estimators	300
min_samples_leaf	1
min_samples_split	2

2.4.3.1. Random Forest

Random Forest 모델에서 max_depth와 n_estimators는 의사결정 Tree의 최대 깊이와 모델에서 사용할 결정 Tree의 총 개수를 설정하였다. 또한 Tree의 내부 노드 분할에 필요한 최소 Sample 수를 지정하는 min_samples_split, 노드에 개별 leaf의 최소 Sample 수를 지정하는 min_samples_leaf를 사용하였다.

RF의 최적화 하이퍼파라미터와 측정기 내부 요인을 추가하여 도출한 모델의 하이퍼파라미터를 Table 3과 Table 4에 각각 나타내었다.

2.4.3.2. LightGBM

LightGBM 모델에서도 RF와 같은 방법을 사용한 결과 하이퍼파라미터와 측정기 내부 요인을 추가하여 도출한 모델의 하이퍼파라미터를 Table 5와 Table 6에 나타내었다.

2.4.3.3. XGBoost

XGBoost 모델에서는 subsample로 훈련 데이터의 일부를 무작위로 선택하여 모델을 훈련하는 비율을 지정하고, L1 및 L2 정규화 항을 조절하는 reg_lambda와 reg_alpha를 사용하였다. min_child_weight는 각 결정 Tree 노드에서 필요한 최소 가중치를 지정하며, max_depth는 결정 Tree의 최대 깊이를 결정한다.

learning_rate는 모델이 각 반복에서 얼마나 빨리 학습할지를 조절하고, gamma는 Tree 노드를 분할하는 데 필요한 최소 손실 감소를 지정한다. 그리고 colsample_bytree

Table 5. LightGBM hyperparameter-1

Parameter	Optimization value
max_depth	6
n_estimators	300
min_samples_leaf	4
min_samples_split	2

Table 6. LightGBM hyperparameter-2

Parameter	Optimization value
max_depth	6
n_estimators	300
min_samples_leaf	4
min_samples_split	2

Table 7. XGBoost hyperparameter-1

Parameter	Optimization value
n_estimators	200
learning_rate	0.2
max_depth	6
min_child_weight	3
subsample	0.8
colsample_bytree	0.9
gamma	1
reg_alpha	1
reg_rambda	1

는 각 결정 Tree를 훈련할 때 사용할 특성의 비율을 나타낸다.

XGBoost 모델의 하이퍼파라미터와 측정기 내부 요인을 추가하여 도출한 모델의 하이퍼파라미터를 Table 7과 Table 8에 나타내었다.

3. 결 과

기상자료개방포털의 기상자료 기온(°C), 누적강수량(mm), 현지기압(hPa), 해면기압(hPa), 풍향(deg), 풍속(m s^{-1}), 습도(%) 등의 데이터 및 본 연구를 위해 개발한 공간선량률 측정 장비의 방사선량률(nSv h^{-1}) 데이터를 활용하여 다중선형회귀, Random Forest, XGBoost,

Table 8. XGBoost hyperparameter-2

Parameter	Optimization value
n_estimators	300
learning_rate	0.1
max_depth	6
min_child_weight	2
subsample	0.9
colsample_bytree	0.7
gamma	0
reg_alpha	1
reg_rambda	1

Table 9. Evaluation for each model

Model	RMSE	R-Squared
MLR-1	43.76	0.01
MLR-3	17.20	0.84
RF-1	36.93	0.29
RF-2	17.31	0.84
RF-3	**2.90**	**0.99**
LightGBM-1	25.06	0.65
LightGBM-2	23.95	0.70
LightGBM-3	**2.39**	**0.99**
XGBoost-1	23.62	0.71
XGBoost-2	22.92	0.73
XGBoost-3	**2.39**	**0.99**

LightGBM 네 가지 모델을 통해 방사선량률(nSv h^{-1})을 예측하고 RMSE (평균제곱근오차) 및 R-Squared (결정계수)를 통해 각 모델의 성능을 평가하여 Table 9에 나타내었다.

Table 9에서 각 모델별 최적화와 측정기 내부 요인을 추가하지 않은 성능 평가결과를 1에 나타냈고, 2는 최적화를 수행한 모델의 결과, 3은 최적화와 측정기 내부 요인을 추가한 후의 성능의 결과를 나타냈다.

기상관측 데이터와 방사선량률(nSv h^{-1})과의 상관계수를 토대로 변수를 선정하여 기계학습 모델별 최적화 결과, LightGBM 모델은 RMSE 23.95 및 R-Squared 0.70으로, XGBoost 모델은 RMSE 22.92 및 R-Squared 0.73의 성능으로 최적화되었다.

GridSearch 방법을 통해 하이퍼파라미터(Hyperparmeter)를 최적화하고 측정기 내부 변수를 추가하여 평가

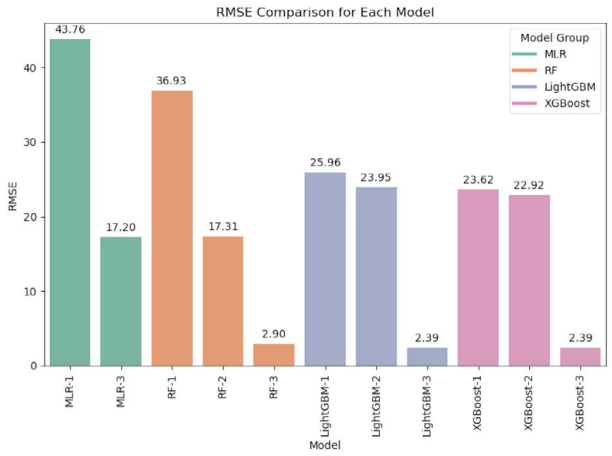

Fig. 3. RMSE for each model.

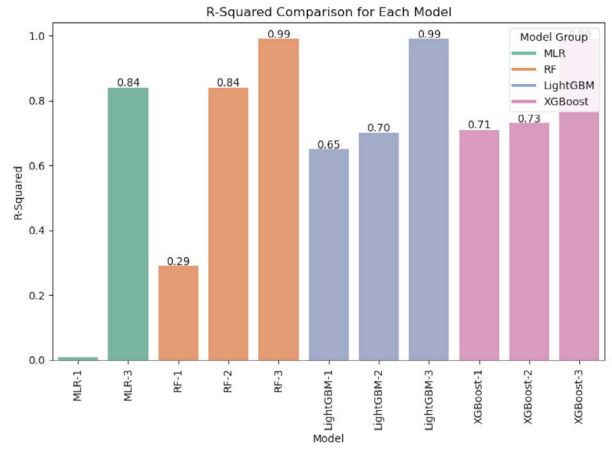

Fig. 4. R-squared for each model.

한 RMSE 결과는 LightGBM과 XGBoost 모델이 2.39로 가장 좋은 성능을 나타냈고, Random Forest는 2.9로 좋은 성능을 나타낸 것을 확인하였다.

이중 다중선형회귀의 경우, 다른 모델과 달리 하이퍼파라미터가 없거나 제한적이기에 제외하였고, 결과가 17.2로 성능이 낮게 도출되어 다중선형회귀 모델인 MLR (Multiple Linear Regression, 다중선형회귀)은 방사선량률 (nSv h^{-1})을 예측하는 데 있어 성능이 떨어지는 것을 확인하였다. 그리고 R-Squared 결과는 RF (Random Forest), LightGBM, XGBoost 모두 0.99로 모델을 잘 설명할 수 있는 좋은 성능임을 확인하였다.

이와 관련하여 Table 9에 제시한 각 모델별 RMSE, R-Squared 결과를 Fig. 3과 Fig. 4에 각각 시각화하여 나타내었다.

4. 결 론

최근 방사선/능은 사회적 관심사로 중요한 이슈가 되어 방사선량률 (nSv h^{-1})의 정확한 예측 필요성이 증가하고 있다.

본 연구에서는 방사선에 대한 예측 모델을 개발하고 최적화하기 위해 XGBoost 및 LightGBM과 같은 기계학습 알고리즘을 활용하였고, 전라남도 장성 지역의 6개월 동안의 종관기상관측 데이터와 방사선량률을 토대로 기후조건과 방사선량률 간의 상관관계를 분석하였으며, 상관계수를 토대로 변수를 선정하여 기계학습 모델을 통해 예측하였다. 이러한 결과는 기계학습 기반 모델이 방사선/능의 정확한 예측에 효과적이며, 방사선/능 측정 장비 내부 요인을 입력 변수로 추가함으로써 모델의 정확도를 높일 수 있음을 확인하였다.

국내의 경우 지리적으로 중국 원전과 매우 인접해 있고, 일본과 같이 중국의 원전 사고에 의해 방사선 사고가 발생할 경우 이러한 지리적 여건으로 방사선/능은 사회적으로 환경 및 먹거리에 대한 국민 불안감 형성되어 있다. 이에 대비하여 머신러닝을 기반으로 한 방사선 예측 모델을 개발하고 최적화하는 것이 중요하다. 머신러닝을 활용한 방사선 모니터링은 더 빠르고 정확한 데이터 분석을 가능케 하여 사고 발생 시 즉각적인 대응을 할 수 있다. 국내 원전에서의 방사선 사고에 대비하여 머신러닝 기술을 활용한 예측 모델을 개발하면, 국민들에게 사고 상황에 대한 더 신속하고 정확한 정보 제공이 가능해질 것이다.

이러한 과학 기술의 발전은 국민들에게 더 안전한 환경을 제공함으로써, 원전 사고에 대한 우려를 최소화할 수 있다. 따라서 국내 원전에서의 방사선 안전에 중점을 둔 머신러닝 기반 모델의 연구와 적용은 국가 안전과 국민 건강을 보호하기 위한 필수적인 노력으로 자리잡게 될 것으로 기대된다.

사 사

본 연구는 한국연구재단에서 주관하는 원자력기초연구지원사업의 지원을 받아 수행한 연구과제입니다 (No. 2022M2D2A201634122).

참고문헌

1. Weather data from Jangseong, Jeollanam-do, Korea Meteorological Administration (Feb. 2022 - Mar. 2023).
2. Airborne dose rate (nSv h^{-1}) data, RMTEC (Feb. 2022 - Mar. 2023).
3. Lee S, Environmental factors and meteorological variables inside a radiation meter, Chosun University.
4. Breiman, L. 2001. Random Forests. *Machine Learning* **45**:5-32. https://doi.org/10.1023/A:1010933404324.
5. Chen T and Guestrin C. KDD's16. 2016. Proceeding of 22nd ACM SICKDD International Conference on Knowledge Discovey and Data Mibing. 12 August 2016(785-794). XGBoost: A scalable tree boosting system. https://doi.org/10.1145/2939672.2939785.
6. Ke G, Meng Q and Finley T. 2017. NIPS': 17 Proceedings of the 31st International Conference on Neural Information Processing Systems December 2017 Pages 3149-3157. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. https://doi.org/10.5555/3294996.3295074.