Journal of the Korean Geo-Environmental Society
23(8): 5～10. (August 2022) http://www.kges.or.kr

ISSN 1598-0820 (Print)
ISSN 2714-1233 (Online)
DOI https://doi.org/10.14481/jkges.2022.23.8.5

지하매설물 속성을 활용한 기계학습 기반 지반함몰 위험도 예측모델 개발

Development of Machine Learning Model to Predict the Ground Subsidence Risk Grade According to the Characteristics of Underground Facility

이 성 열[1] · 강 재 모† · 김 진 영[2]
Sungyeol Lee · Jaemo Kang · Jinyoung Kim

Received: May 20th, 2022; Revised: May 23rd, 2022; Accepted: June 14th, 2022

ABSTRACT : Ground Subsidence has been continuously occurring in densely populated downtown. The main cause of ground subsidence is the damaged underground facility like sewer. Currently, ground subsidence is being dealt with by discovering cavities in ground using GPR. However, this consumes large amount of manpower and cost, so it is necessary to predict hazardous area for efficient operation of GPR. In this study, ○○city is divided into 500 m×500 m grids. Then, data set was constructed using the characteristics of the underground facility and ground subsidence in grids. Data set used to machine learning model for ground subsidence risk grade prediction. The purposed model would be used to present a ground subsidence risk map of target area.

Keywords : Ground Subsidence, Sewer, Machine Learning, Ground subsidence prediction model, Ground subsidence risk map

요 지 : 인구 밀집도가 높은 도시 중심지에서 발생하는 지반함몰의 주요 원인은 하수관 및 상수관과 같은 지하매설물의 손상으로 알려져 있다. 이와 관련하여 지반함몰의 원인 규명과 지반함몰 위험 예측에 관한 연구가 꾸준히 수행되고 있다. 현재 지반함몰은 지중탐사레이더를 통해 선제적으로 공동을 발견하여 대응하고 있으나, 이는 인력 및 비용의 소비가 크기 때문에 효율적인 장비의 운영을 위해 위험지역을 예측하고 예측된 지역을 우선순위로 탐사해야 할 필요가 있다. 따라서 본 연구에서는 ○○시의 2개 구를 500m×500m 크기의 그리드로 분할하고, 해당 그리드 내의 지하매설관 속성과 지반함몰 발생 데이터를 활용하여 데이터셋을 구축하였다. 구축된 데이터셋으로 기계학습을 통한 적절한 지반함몰 위험등급 예측 모델을 제시하였고, 제시된 모델을 활용하여 대상지역의 지반함몰 위험지도를 제시하고자 하였다.

주요어 : 지반함몰, 지하매설물, 기계학습, 지반함몰 위험지도, 지반함몰 위험 예측모델

1. 서 론

인구 밀집도가 높은 도심지를 중심으로 지반함몰이 꾸준하게 발생하고 있어 많은 시민들이 불안에 떨고 있다. 지반함몰의 발생은 인명 및 재산피해, 사회 인프라 마비 등을 야기하므로 사고 원인 규명 및 대비가 필요하다. 지반함몰의 주요 원인으로는 하수관 및 상수관로의 손상, 인접 굴착공사 등으로 꼽히고 있으나(서울시, 2016), 복합적인 원인에 의해 발생하는 현상이기 때문에 명확한 원인을 찾아 사고를 선제적으로 대비하는 것은 매우 어려운 일이다.

지반함몰의 발생은 지하매설물의 손상으로 인해 지반 내 공동이 발생하며 발생된 공동이 확장되어 지반 상부가 붕괴되는 메커니즘을 보이고 있으므로(Kim et al., 2017), 지반 내 공동을 탐사하는 지표투과레이더(GPR, Ground Penetrating Radar)를 활용하여 공동을 사전에 발견하여 사고에 대비하고 있다. 하지만, 지표투과레이더는 인력 및 시간, 비용의 소비가 매우 크기 때문에 비효율적이므로 최적의 탐사 지역을 선정하기 위한 다양한 지반함몰 위험도 예측 방법과 지반함몰 원인 규명에 관한 연구가 제안되고 있다. Kuwano(2006) 등은 일본의 표준사를 활용하여 지반 내 공동 발생 메커니즘을 모형실험을 통해 규명하였고, Mokunoki(2009)는 실내 모형실험을 통하여 하수관 균열 시 발생하는 지반 내 공동 발생 메커니즘을 규명하고 X-ray와 CT로 시각화하는 연구를 수행하였다.

국내의 지반함몰 발생의 주요원인은 지하매설물의 손상(86.2%)으로 조사되었으며(Seoul City, 2014), Jin(2018)은 AHP

1) Postdoctoral Researcher, Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology
† Senior Researcher, Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology (Corresponding Author : jmkang@kict.re.kr)
2) Senior Researcher, Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology

분석을 통한 지반함몰 발생 영향인자의 가중치 분석을 실시하였으며, 그 결과 지하매설물의 노후화가 가장 큰 영향인자로 선정되었다. Han(2017)은 하수관로의 CCTV 자료와 GPR의 조사 자료를 활용하여 지반함몰 위험도 평가를 작성하였고, Kim(2018)은 기계학습의 알고리즘인 로지스틱 회귀분석 모델을 활용하여 지반함몰 위험도 산정식을 제안하였다. 또한, Lee(2022) 등은 기계학습 알고리즘을 활용해 지반함몰 발생 예측 모델을 제안하고 비교하는 연구를 발표하였다. 이와 같이 국내를 중심으로 다양한 기법을 통한 지반함몰 위험도 예측의 연구가 꾸준히 수행되고 있으나, 실질적으로 활용할 수 있는 대상 지역의 지반함몰 위험도를 제시할 수 있는 모델은 아직 미흡한 실정이다.

따라서 본 연구에서는 ○○시 ○○구의 지하매설물 속성정보 중 매설년수, 관경, 관로 길이를 활용하여 데이터셋을 구축하였으며, 구축된 데이터셋을 기계학습 알고리즘 중 분류 문제 해결에 특화된 랜덤포레스트(Random Forest, RF), XGBoost, LightGBM에 적용하여 지반함몰 위험도 예측 모델을 선정하고, 선정된 모델을 통해 지반함몰 위험지도를 제시하고자 하였다.

2. 기계학습 알고리즘

2.1 랜덤 포레스트(RF)

RF 알고리즘은 다수의 Tree 기반의 모델을 활용하여 결과를 도출하는 배깅(Bagging) 기법 중 하나로 하나의 Tree를 활용하여 결과를 도출하는 Decision Tree 알고리즘에서 발전된 모델이다. RF 모델은 데이터 간의 상관성이 높지 않을 경우에도 대상을 일반화하여 분류하는데 탁월한 모델로 알려져 있으며(Lee, 2022), Fig. 1과 같이 다수의 Tree에서 도출된 가중치를 합치거나 최적의 가중치를 채택하는 방법으로 결과를 도출한다.

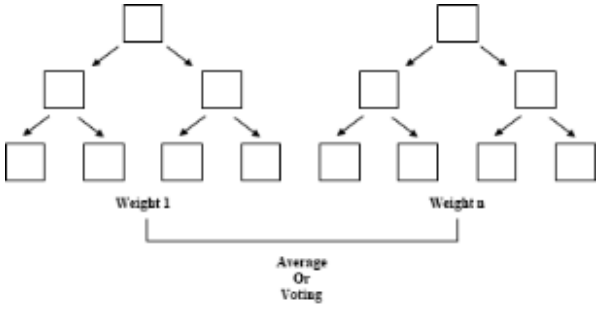


Fig. 1. Conceptual diagram of RF

2.2 XGBoost

XGBoost 알고리즘은 단일 모델을 순차적으로 적용하여 모든 모델이 가중치 선정에 반영되는 부스팅(Boosting) 기법 중 하나의 모델이며, 경사 하강법을 적용하여 모델의 가중치를 산정하는 Gradient Boosting 모델이다(Lee, 2022). XGBoost는 선형과 트리기반으로 과적합에 대한 위험이 상대적으로 적으며, 모델의 처리 속도가 빠르고 성능이 우수하여 빅데이터를 적용하기 알맞은 모델이다(Ha et al., 2017). XGBoost 모델에서 결정해야 할 대표적인 Hyperparmeter는 트리의 개수, 트리의 깊이 등이 있다.

2.3 LightGBM(Light Gradient Boosting Machine)

LightGBM 알고리즘은 XGBoost와 같은 부스팅 기법 중 GBM(Gradient Boosting Machine)의 하나이며, Tree를 순차적으로 적용하여 최종 가중치를 도출하는 모델이다. GBM은 우수한 성능으로 분류 및 회귀 문제 해결에 주로 사용되고 있으나, 모델의 처리 속도가 느리고 메모리의 효율이 떨어지는 단점을 갖고 있다(Lee et al., 2022). 하지만 LightGBM은 Fig. 2와 같이, Tree 중 일부만 빠르게 계산하여 가중치를 도출함으로 모델의 처리 속도를 향상시켜 GBM의 단점을 개선하였다(Lee et al., 2022).

2.4 모델의 평가지표

본 연구에서는 기계학습에 데이터셋을 학습시킨 모델의 성능을 평가하기 위한 지표로 일반적으로 분류 모델의 평가지표로 활용되고 있는 정확도(Accuracy)와 F1-Score, AUC(Area Under the Curve)를 선정하였다.

정확도는 학습 데이터(Train)와 평가 데이터(Test)의 정확도를 도출하여 과적합 여부를 확인하는데 사용되었으며, 실제 지반함몰 위험도를 정확히 구분한 정도를 의미하는 지표이다.

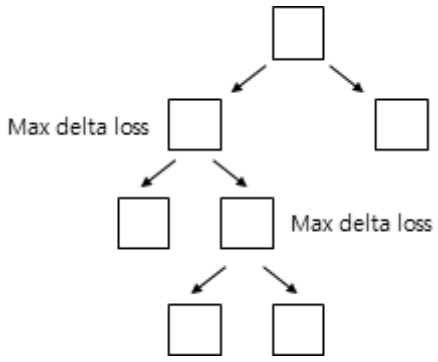


Fig. 2. Conceptual diagram of LightGBM

F1-Score는 본 연구에서 활용된 데이터셋과 같은 불균형 데이터에서 예측 모델이 True라고 예측한 데이터 중 실제 True인 데이터의 수(Precision)와 실제 True인 데이터에서 True라고 예측한 데이터의 수(Recall)의 조화 평균을 나타낸 지표이다. 이를 활용하면 모델이 각각의 Class를 올바르게 분류했는지에 대한 평가가 가능하다.

AUC는 Table 1과 같은 기준으로 모델의 성능을 평가할 수 있는 지표(Fawcett, 2005)로 Recall과 Specificity를 이용하여 나타낸 ROC(Receiver Operating Characteristic) Curve에 면적을 의미한다. Eq. (1)~(5)는 평가지표의 산출 방법을 나타낸 식이며, Table 1은 모델의 분류 결과를 나타낸 Confusion Matrix, Fig. 3은 ROC Curve의 모식도이다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Recall\,(Sensitivity) = \frac{TP}{TP+FN} \quad (2)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$F1\,Score = 2\times\frac{\mathrm{Precision}\times Recall}{\mathrm{Precision}+Recall} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

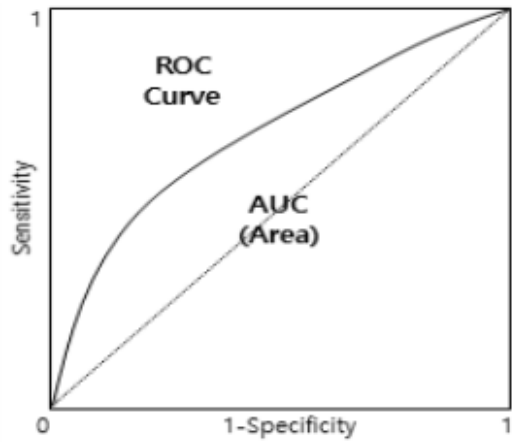


Fig. 3. Conceptual diagram of ROC curve

Table 1. Confusion matrix

Confusion matrix		Prediction	
		Negative	Positive
Reference	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

3. 데이터셋 구축

3.1 대상지역의 구분

○○시의 ○○구와 △△구의 지하매설물 속성 정보와 지반함몰 발생정보를 활용하여 지반함몰 위험도 예측 모델을 학습하고 결과를 평가하였다. 데이터셋의 구축을 위해 ArcGIS 프로그램을 활용하여 Fig. 4와 같이 해당 구를 500m×500m 크기의 160개 그리드(Grid)로 구분하였으며 각 그리드 내에 속해있는 지하매설물의 속성 값을 추출하고, 그리드 내에서 발생한 지반함몰 발생 개수를 파악하여 개수에 따른 지반함몰 위험등급을 임의로 구분하였다.

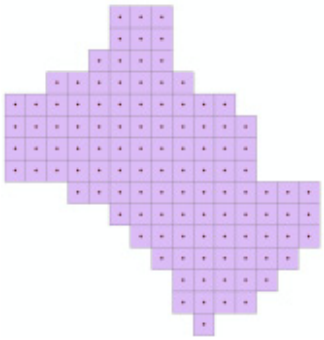


Fig. 4. Grid division of area

3.2 지하매설물 속성 데이터

데이터셋 구축에 활용된 지하매설물의 종류는 하수, 상수, 전기, 가스, 난방, 통신관이며, 6종의 지하매설물을 1개의 관으로 고려하여 속성정보를 추출하였다. 각 그리드 내에 속해있는 지하매설물에 대한 속성정보 중 결측값이 적어 분석에 활용이 용이한 영향인자는 관의 활용년수와 관의 직경, 관로의 길이이며, 활용년수 및 관의 직경에 대한 단위로 관로의 길이를 사용하였다. 예를 들어, 대상지역의 지하매설물 중 활용년수가 1년차인 전체관로의 길이를 종합하는 방법으로 데이터셋을 구축하였고, 활용년수는 1년, 관경은 100m를 범주로 관로의 길이를 종합하였다.

3.3 지반함몰 위험등급

데이터셋은 전술한 지하매설물 속성(활용년수, 관경) 별 관로의 길이와 지반함몰 위험등급으로 구성하였다. 지반함몰 위험등급은 그리드 내에 발생한 지반함몰 개수를 파악한 뒤, 모델이 최적의 분류 성능을 도출할 수 있도록 임의로 3단계와 4단계로 구분하였다.

지반함몰 발생 개수는 1개 그리드 내에 0~44개의 범위로

나타났으며, 발생 개수의 대다수는 0개에 집중되어 있다. 위험등급 3단계는 지반함몰 개수가 0개, 1~10개, 11개 이상으로 구분하였으며, 4단계는 0개, 1~10개, 11~20개, 20개 이상으로 구분하였다.

3.4 구축된 데이터셋의 조건

지하매설물 속성 중 Table 2와 같이 각각의 범주(활용년수 1년, 관경 100m)에 대한 관로 길이의 데이터와 지반함몰 위험등급의 데이터를 구축하고 최적의 모델 성능을 발휘하는 데이터셋을 탐색하고자 4개 조합의 데이터셋을 구축하였다.

활용년수의 경우, 5년과 10년을 단위로 데이터를 병합하였으며, 관경은 100m를 단위로 병합하였다. 또한, 지반함몰 위험등급은 3단계와 4단계로 구분하여 데이터셋을 적용하였다. Table 3은 데이터셋 조합의 구성을 나타낸 표이며, Year은 활용년수를 의미하고, Dimeter에서의 100은 관경의 100mm 단위를 의미한다.

Table 2. Category of factors

Factors		Category
Year (year)	5	1~5, 6~10, 11~15, 16~20, 21~25, 26~30, 31~35, 36~40, 41~45, 46~50
	10	1~10, 11~20, 21~30, 31~40, 41~50
Diameter (mm)		1~100, 101~200, 201~300, 301~400, 401~500, 501~600, 601~700, 701~800, 801~900

Table 3. Configuration of datasets

	Year (year)	Diameter (mm)	Risk level
M-1	5	100	3
M-2	5	100	4
M-3	10	100	3
M-4	10	100	4

4. 기계학습 모델의 적용 방법 및 결과

4.1 기계학습 모델의 적용 방법

구축된 데이터셋을 기계학습 알고리즘 중 분류문제 해결에 우수한 성능을 발휘하는 RF, XGBoost, LightGBM에 적용하여 지반함몰 위험도 예측 모델을 제시하고, 그 결과를 비교하여 최적의 모델을 선정하고자 하였다. 이를 위해 Python 3.8과 Scikit-learn 라이브러리를 활용하였으며, 모델의 과적합을 회피하기 위해 교차검증 알고리즘인 StratifiedKFold을 적용하였으며 데이터를 학습(Train) 데이터와 평가(Test) 데이터를 80:20의 비율로 분할하여 모델에 적용하였다.

데이터셋을 적용한 모델이 최적의 결과를 도출하고 과적합(OverFitting)을 회피할 수 있도록 Hyper-parameter를 튜닝(Tuning) 하였다. 튜닝의 과정은 시행착오법으로 최적의 결과가 도출되는 Hyper-parameter를 결정하였으며, Table 4는 모델에 적용된 주요 Hyper-parameter를 나타낸 표이다. RF 모델의 주요 Hyper-parameter의 범위는 모델의 개수를 의미하는 estimators(500), Tree의 깊이를 의미하는 Max depth(2)이고, XGB(XGBoost) 모델의 주요 변수의 범위는 estimators (300), Max depth(1), learning rate(0.01), LightGBM은 estimators(500), Max depth(1), learning rate(0.001)이며, learning rate는 모델의 학습률을 의미한다.

Table 4. Summary of hyper parameters in the model

Model	Hyper parameter
RF	Estimators (500) Max depth (2)
XGB	Estimators (300) Max depth (1) Learning rate (0.01)
LightGBM	Estimators (500) Max depth (1) Learning rate (0.001)

4.2 기계학습 적용 결과 및 모델 선정

최적의 지반함몰 위험도 예측 모델을 제시하고자 구축된 데이터셋을 Hyper-parameter가 조정된 모델에 적용한 결과를 Table 5에 나타냈고, F1-Score의 micro는 전체 클래스의 F1-score의 평균 값을 의미한다.

모델의 평가지표를 비교한 결과, Test Score와 F-1 Score 모두

Table 5. Result comparison for each models

Model		Test score	Train score	F-1 score (micro)
M-1	RF	0.688	0.705	0.50
	XGB	0.750	0.783	0.72
	LightGBM	0.719	0.758	0.66
M-2	RF	0.744	0.798	0.39
	XGB	0.744	0.76	0.39
	LightGBM	0.750	0.797	0.40
M-3	RF	0.775	0.817	0.75
	XGB	0.788	0.839	0.77
	LightGBM	0.644	0.709	0.51
M-4	RF	0.756	0.808	0.40
	XGB	0.769	0.887	0.41
	LightGBM	0.781	0.861	0.45

M-3의 XGB에서 가장 우수한 것으로 나타났으며, 전반적인 모델의 성능도 XGB가 우수하게 나타났다. 또한, M-1과 M-3, M-2와 M-4의 비교를 통해 활용년수를 10년으로 구분한 것이 모델의 성능이 우수한 것으로 나타났다. M-1과 M-2, M-3과 M-4를 비교하면 위험등급에 따른 모델의 평가지표를 비교할 수 있으며, 3단계로 위험등급을 구분하는 것이 모델의 성능이 월등히 우수한 것으로 나타났다.

따라서 지반함몰 위험등급 예측에 가장 적합한 모델은 M-3의 XGB로 선정하였다. 선정된 모델의 Test와 Train Score의 차이는 0.051로 과적합을 회피한 것으로 판단되며, F-1 Score도 0.77로 나타나 모델이 각 Class의 분류 성능도 우수한 것으로 판단된다. 해당 모델의 성능을 검증하기 위한 AUC를 알아보기 위해 Fig. 5와 같이, ROC Curve를 확인하였다. 그 결과, 각 Class의 AUC가 모두 0.9 이상으로 모델의 성능이 우수하다고 평가할 수 있는 기준인 0.8을 상회하는 것으로 나타났다.

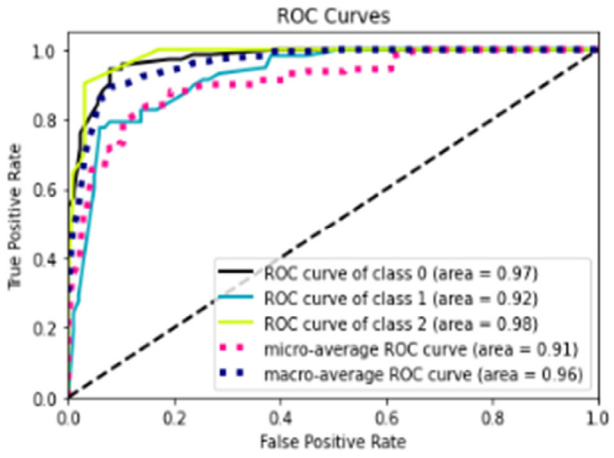
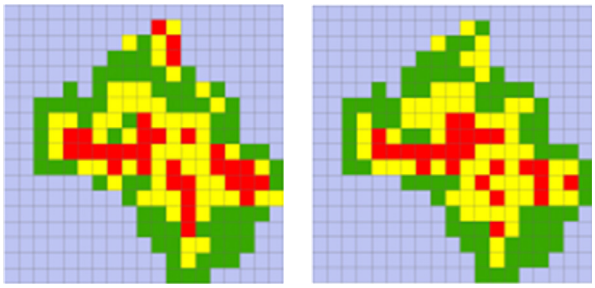


Fig. 5. ROC curve of M-3 XGB model

4.3 선정 모델을 활용한 위험지도

본 연구에서 제시된 M-3 XGB 모델에서 예측한 지반함몰 위험등급을 연구 대상지역에 적용하여, 실제 위험도와 모델이 예측한 위험도를 비교하였다. Fig. 6(a)는 실제 위험도를 나타낸 그림이며, Fig. 6(b)는 모델이 예측한 위험도를 나타낸 그림이다. 위험지도에서 초록색은 위험도 '0'을 의미하며, 노란색은 '1', 빨간색은 '2'를 나타낸다. 대상지역의 서부지역에서는 모델이 지반함몰 위험도를 상대적으로 잘 예측하였으나, 북부지역의 경우 위험도를 실제보다 낮게 예측한 경향이 나타났다. 이는 북부지역의 지하매설물의 밀집도가 상대적으로 낮아 발생한 결과로 판단되며, 추후 밀집도를 영향 인자로 추가하여 지반함몰 예측 모델을 개발해야 할 필요가 있을 것으로 판단된다.



(a) Reference (b) Prediction

Fig. 6. Ground subsidence risk map

5. 결 론

본 연구는 ○○시의 2개구를 대상으로 ArcGIS 프로그램을 활용하여 500m×500m 크기의 그리드 분할 후 그리드의 지반함몰 위험도 예측을 위한 기계학습 모델을 제시하고자 그리드 내에 발생한 지반함몰 개수를 통해 위험등급을 산정하였고, 지하매설물의 속성인 활용년수와 관의 직경을 관로길이로 계산하여 데이터셋을 구축하였다. 구축된 데이터셋을 RF, XGB, LightGBM에 학습하여 최적의 모델을 선정하였으며, 선정된 모델을 통해 대상지역의 지반함몰 위험지도를 작성하였다. 본 연구의 요약과 결론은 다음과 같다.

(1) 지반함몰 위험도 예측 모델을 위해 다양한 조건에서의 데이터셋을 구축하여 평가지표를 비교한 결과, 활용년수는 10년, 위험등급은 3단계로 구분할 경우 예측 모델의 평가지표가 우수하게 나타났다.
(2) 다양한 조건의 데이터셋을 기계학습 알고리즘(RF, XGB, LightGBM)에 적용한 결과, XGB 모델(M-3)에서 상대적으로 높은 정확도와 F-1 Score가 도출되었고, AUC도 0.8을 상회하여 해당 모델을 가장 적절한 모델로 선정하였다.
(3) 선정된 모델(M-3 XGB)에서 예측된 지반함몰 위험도를 통해 ArcGIS를 활용한 위험지도의 작성이 가능하며, 실제 발생한 데이터를 토대로 작성된 위험지도와 비교 시 유사한 형태의 지반함몰 위험지도가 도출되었다.

본 연구를 통해 제안된 지반함몰 위험도 예측 모델 및 위험지도를 통해 복잡한 원인으로 발생하는 지반함몰 위험도의 예측이 가능할 것으로 판단되며, 지중탐사레이더와 같은 지반함몰 사고 방지 대응 시 우선순위 지역을 선정하여 효율적인 탐사가 이루어 질 것으로 기대된다.

감사의 글

본 연구는 (22주요-대1-임무)지하 공간 정보 정확도 개선 및 매설관 안전관리 기술개발 (3/3) 지원으로 수행되었으며, 이에 깊은 감사를 드립니다.

References

1. Breiman, L. (2001), Random Forest. Machine Leaerning, Kluwer Academic Publishers, 45, pp. 5~32.
2. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Liu. (2017), LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Part of Advances in Neural Information Processing Systems 30.
3. Ha, J. E., Shin, H. C. and Lee, Z. K. (2017), Korean text classification using randomforest and XGBoost focusing on Seoul metropolitan civil complaint data, The Journal of Bigdata, Vol. 2, Issue 2, pp. 95~104.
4. Han, M. S. (2017), A risk assessment of ground subsidence by GPR and CCTV investigation, Master's thesis, Seoul National University of Science and Technology (In Korean).
5. Jin, Y. S. (2020), The Analysis on Correlation of Precipitation and Risk Factors to the Soil Subsidence, Ph D. dissertation, Chonnam National University, pp. 104~105 (In Korean).
6. Kim, K. Y. (2018), Susceptibility Model for Sinkholes Caused by Damaged Sewer Pipes Based on Logistic Regression, Master's thesis, Seoul National University.
7. Kuwano, R., Horii, T., Kohashi, H. and Yamauchi, K. (2006), Defects of sewer pipes causing cave-in's in the road, Proc. 5th International Symposium on New Technologies for Urban Safety of Mega Cities in Asia,Phuket, Thailand, pp. 347~353.
8. Lee, S. Y., Kim, J. Y., K, J. M. and Baek, W. J. (2022), Comparison of machine learning models to predict the occurrence of ground subsidence according to the characteristics of sewer, Journal of Korean Geo-Environmental Society, Vol. 23, Issue 4, pp. 5~10.
9. Mukunoki, T., Kuwano, N., Otani, J. and Kuwano, R. (2009), Visualization of three dimensional failure in sand due to water inflow and soil drainage from defected underground pipe using X-ray CT, Soils and Foundations, Vol. 49, No. 6.
10. Seoul Seokchon-dong Cavity Cause Investigation Committee, (2014), Cause Analysis of Cavity at Seokchon Underground Roadway and Road Cavity.
11. Seoul Institute (2016), The Road Subsidence Conditions and Safety Improvement Plans in Seoul (In Korean).
12. Tom Fawcett (2005), An introduction to ROC analysis, Patter Recognition Letters, Edited by Francesco Tortorella, Vol. 27 Issue 8, pp. 861~874.