Journal of the Korean Geo-Environmental Society
23(4): 5～10. (April 2022) http://www.kges.or.kr

ISSN 1598-0820 (Print)
ISSN 2714-1233 (Online)
DOI https://doi.org/10.14481/jkges.2022.23.4.5

하수관로 특성에 따른 지반함몰 발생 예측을 위한 기계학습 모델 비교

Comparison of Machine Learning Models to Predict the Occurrence of Ground Subsidence According to the Characteristics of Sewer

이 성 열[1] · 김 진 영[†] · 강 재 모[2] · 백 원 진[3]
Sungyeol Lee · Jinyoung Kim · Jaemo Kang · Wonjin Baek

Received: January 11th, 2022; Revised: January 20th, 2022; Accepted: March 17th, 2022

ABSTRACT : Recently, ground subsidence has been continuously occurring in downtown areas, threatening the safety of citizens. Various underground facilities such as water and sewage pipelines and communication pipelines are buried under the road. It is reported that the cause of ground subsidence is the deterioration of various facilities and the reckless development of the underground. In particular, it is known that the biggest cause of ground subsidence is the aging of sewage pipelines. As an existing study related to this, several representative factors of sewage pipelines were selected and a study to predict the risk of ground subsidence through statistical analysis has been conducted. In this study, a data SET was constructed using the characteristics of OO city's sewage pipe characteristics and ground subsidence data, The data set constructed from the characteristics of the sewage pipe of OO city and the location of the ground subsidence was used. The goal of this study was to present a classification model for the occurrence of ground subsidence according to the characteristics of sewage pipes through machine learning. In addition, the importance of each sewage pipe characteristic affecting the ground subsidence was calculated.

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

요 지 : 최근 도심지에서는 지반침하가 지속적으로 발생하여 시민의 안전을 위협하고 있다. 상하수도관, 통신관 등 각종 지하시설물이 도로 밑에 매설되어 있다. 지반침하의 원인으로는 도심지에 매설되어 있는 각종 시설물의 노후화와 급격한 도시화로 인한 지하 난개발로 인한 것으로 보고되고 있다. 특히 지반침하의 가장 큰 원인은 하수관로의 노후화로 알려져 있다. 이와 관련된 기존 연구로는 하수관로의 대표적인 몇 가지 요인을 선정하여 통계분석을 통해 지반침하 위험을 예측하는 연구가 진행되었다. 본 연구에서는 OO시의 하수관 특성과 지반침하 데이터를 이용하여 데이터셋을 구축하고, OO시의 하수관 특성과 지반함몰 발생 위치 데이터로 구축된 데이터셋으로 기계학습을 통한 하수관 특성에 따른 지반함몰 발생 분류 모델들을 비교하여 적절한 모델을 선정하고자 하였으며, 선정된 모델에서 도출된 지반함몰에 영향을 미치는 하수관 특성별 중요도를 산정하고자 하였다.

주요어 : 지반함몰, 하수관로, 기계학습, 지반함몰 예측 모델

1. 서 론

최근 고양시 마두역 주변에서 발생한 지반함몰과 같이 지반함몰이 꾸준하게 발생하고 있어 시민들을 불안에 떨게 하고 있다. 지반함몰은 여러 가지 복합적인 원인으로 발생하지만 서울시(2016)에서 발표한 지반침하 현상의 원인을 살펴보면 하수관로 손상(84.5%), 상수관로 손상(1.7%), 인접 굴착 공사 등 기타(13.8%)로 하수관로가 가장 큰 영향을 미친다고 발표하였다. 지반함몰 발생 현황을 살펴보면 인구가 밀집되어 있는 도심지를 중심으로 발생하며, 사고 시 재산 피해 및 인명피해를 야기하므로 사고의 예방을 위한 선제적인 조치가 필요하다. 현재 지반함몰 발생의 관리를 위해 지표투과레이더(GPR, Ground Penetrating Radar)를 활용하여 지반 내 공동을 탐사하고 적절한 조치를 취하고 있다. 하지만, 광범위한 지역을 우선순위 없이 탐사하는 것은 매우 비효율적이며, 많은 비용이 소비되기 때문에 복잡한 지반함몰 발생의 원인을 분석하고 위험도를 예측하는 방법이 제안되고 있다. 공동의 발생은 지반 내 매설 구조물과 토립자 재료사이의 이질적 차이로 인해 투수계수가 변동됨에 따라 물길의 생성으로 인한 지반의 강도 저하가 공동 발생 위험성을 증가시키며 여러 연구자들이 모형실험을 통하여 토조에 하수관의 손상을 모사하여 손상에 따른 지반함몰 발생 메커니즘을

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 : goldcamp@kict.re.kr)
2) Senior Researcher, Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology
3) Professor, Department of Rural and Biosystems Engineering, Chonnam National University

발표하였다(Kuwano et al., 2006). 국내의 지반함몰 발생의 가장 주요한 원인으로는 하수관 손상으로 조사 되었으며(Seoul City, 2016), AHP 분석을 통하여 하수관 특성에 따른 지반함몰 발생 영향인자의 가중치를 분석한 결과 노후화에 기인한 하수관의 손상이 가장 큰 가중치를 부여받은 것으로 나타났다(Jin, 2018). 또한, 기계학습의 알고리즘 중 하나인 로지스틱 회귀분석을 통하여 지반함몰 위험도 평가를 수행한 연구에서는 지반함몰에 영향을 미치는 하수관 특성으로 매설 년수, 경사도, 관의 종류, 관의 직경, 관의 길이, 매설심도, 관저고로 선정하였다(Kim et al., 2017).

이와 같이 지반함몰의 원인 분석 및 다양한 기법을 활용한 위험도 예측 방법을 통해 사고 발생을 사전에 대응하기 위한 연구가 꾸준히 진행되고 있다. 기존 연구로는 전문가 의견을 반영한 통계기법, 로지스틱 회귀분석 등을 활용하여 지반함몰 위험도를 예측하는 연구가 많이 진행되어 왔지만 기계학습을 이용하여 지반함몰 위험도를 분석한 연구는 미흡한 실정이다. 기계학습은 컴퓨터 프로그램이 데이터와 처리 경험을 이용한 학습을 통해 정보 처리 능력을 향상시키는 기술로 방대한 양의 데이터 분석에 활용이 용이한 기법이다.

따라서 본 연구에서는 OO시의 하수관 특성과 지반함몰 발생의 데이터를 활용하여 데이터셋을 구축하고 기계학습의 분류 알고리즘인 랜덤포레스트(Random Forest, RF), LightGBM, KNN을 활용하여 하수관 손상에 따른 지반함몰 발생 분류 모델의 평가지표를 비교하여 적절한 모델을 제안하고자 하였다.

2. 기계학습 알고리즘

2.1 랜덤포레스트(RandomForest, RF)

랜덤포레스트는 기계학습 분류 기법 중 다수의 모델을 활용하여 최적의 모델을 결과로 도출하는 기법인 배깅(Bagging)의 대표적인 알고리즘으로, 트리(Tree) 모형 기반인 의사결정나무 모델이 다수 구성되어 있는 특징을 갖고 있는 알고리즘이다(Breiman, 2001). RF는 Fig. 1에 나타난 바와 같이 True or False 방법의 이진 분류를 실시하여 문제의 결과를 도출해내며, 입력 데이터 사이의 상관성이 높지 않은 경우 데이터를 일반화하여 정확도가 높은 결과를 나타내는 특징이 있다.

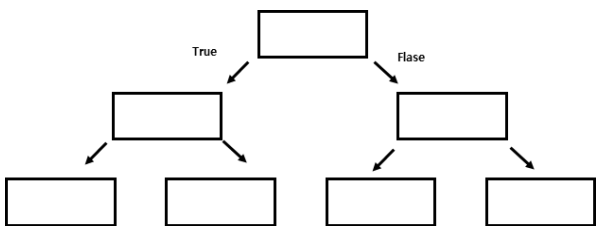


Fig. 1. Conceptual diagram of RF

2.2 LightGBM(Light Gradient Boosting Model)

LightGBM 알고리즘은 의사결정나무 모델을 순차적으로 적용하여 모든 모델이 최종 가중치(Weight)를 산정할 때 개입되는 강한 분류 기법인 Boosting의 종류 중 하나이다. 기존 GBM(Graident Boosting Model)은 분류 예측 시 높은 성능을 나타내어 다양한 분야에서 활용되고 있지만, 빅데이터 적용 시 분석 속도와 사용되는 메모리의 효율 측면에서 비효율적인 단점을 갖고 있다. 따라서 LightGBM 알고리즘은 Fig. 2에 나타낸 바와 같이 트리 모델 중 일부만 빠르게 계산하여 GBM의 장점을 살리고 모델링의 시간을 단축시켜 단점이 해결된 모델로 개발되었다(Lee et al., 2020). 하지만, 리프(Leaf) 중심으로 분할되는 알고리즘 특성상 과적합(Over-Fitting)에 대한 위험이 높은 편이다(Ke et al., 2017).

2.3 KNN(K-Nearest Neighbor)

KNN은 분류 모델에 특화된 알고리즘으로 결론을 도출하고자 하는 데이터를 중심으로 거리를 계산하여 군집하고 인근 데이터의 특성을 통해 범주를 분류하는 직관적인 모델이다. KNN 알고리즘은 Fig. 3과 같이 군집을 위하여 3차원 유클리디안 거리 공식(Euclidean Distance)을 활용하여 데이터

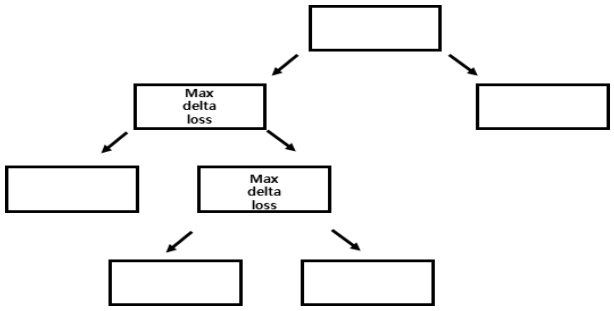


Fig. 2. Conceptual diagram of LightGBM

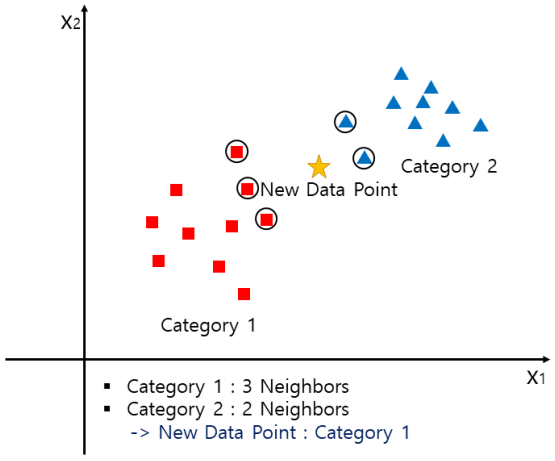


Fig. 3. K-Nearest Neighbor (KNN) Method

사이의 거리를 산정한 뒤 인근 데이터의 개수 K의 다수결 원칙에 따라 의사결정을 수행한다. 알고리즘에서 K를 큰 값으로 설정할 경우 Under-fitting의 위험이 존재하므로 결과 도출에 적절한 K의 개수를 설정하는 것이 알고리즘의 핵심이다.

2.4 분류모델의 평가

본 연구에서는 일반적으로 기계학습 분류 모델의 평가를 위해 일반적으로 사용되는 지표인 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-Score, AUC(Area Under the Curve)로 선정하였다. 정확도는 실제 지반함몰 데이터와 비교하였을 때 모델이 발생여부를 정확히 예측한 비율을 의미하며, 정밀도는 모델이 지반함몰 발생으로 예측한 데이터(TP+FP) 중 실제 지반함몰이 발생한 데이터(TP)의 비율을 나타낸다. 또한, 재현율은 실제 지반함몰이 발생한 데이터(TP+FN) 중 모델이 지반함몰로 예측(TP)한 비율을 의미하고, F1-Score는 정밀도와 재현율을 조화 평균하여 지반함몰 발생 여부에 대해 모델이 분류 성능을 나타내는 지표이다. 모델의 성과 지표를 나타내는 AUC는 재현율과 실제 지반함몰 미발생 데이터를 모델이 미발생으로 예측한 지표인 특이도(Specificity)를 이용해 나타낸 ROC(Receiver Operating Characteristic) Curve의 면적을 의미한다. AUC가 0.8 이상일 시 해당 모델의 성능이 우수하다고 평가할 수 있다(Fawcett, 2005). 각 지표의 산출식을 Eq. (1)~(5)에 나타냈으며, Fig. 4는 ROC Curve의 개념도이다.

위의 평가 지표를 산출하기 위해서는 Table 1에 제시된 Confusion Matrix를 도출해야 하고, Table 1에 나타난 TN은 실제 지반함몰 미발생 데이터 중 모델이 미발생으로 예측한 값, FP는 지반함몰 미발생 데이터 중 모델이 발생으로 예측한 값, FN은 지반함몰 발생 데이터 중 모델이 미발생으로 예측한 값, TP는 지반함몰 발생 데이터 중 지반함몰 발생으로 예측한 값을 의미한다.

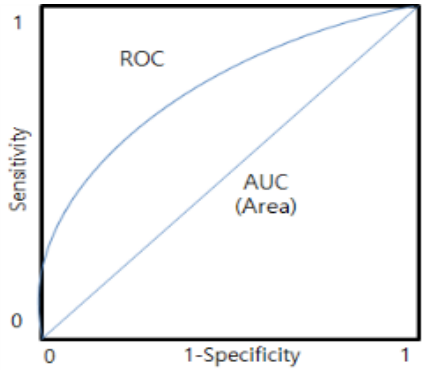


Fig. 4. Conceptual diagram of ROC Curve

Table 1. Confusion Matrix of binary classification

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

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

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

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

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

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

3. 예측모델 데이터의 구성

3.1 데이터셋 구축

본 연구에서는 지반함몰의 주요 원인 시설물로 OO시의 지하 매설관로 중 대상 관로로 하수관로를 선정하였으며, 하수관로 특성에 따른 지반함몰 발생의 분류 모델을 선정하기 위하여 분석을 위한 데이터셋을 구축하였다. 데이터 구축 방법으로는 ArcGIS Near Tool을 활용해 2008년부터 2014년까지 기록된 지반함몰 발생 위치에서 가장 인접한 하수관로를 추출하고, 하수관로 속성데이터와 지반함몰 속성데이터를 결합하여 데이터 분석을 위한 하나의 데이터셋으로 구축하였다. 기존 연구에 따르면 지반함몰의 주요 원인으로는 하수관의 손상이 대부분인 것으로 나타나 있으며(Seoul Metropolitan City, 2017), 기초 분석 결과 OO시 대부분의 하수관은 HP관이 매설되어 있기 때문에 원활한 분석을 위해 일률적 데이터 특성을 갖는 원형관인 HP관을 중심으로 데이터셋을 구성하였다. 데이터셋 구축을 위하여 ArcGis의 Tool을 활용해 Fig. 5와 같이 지반함몰이 발생한 지점에서 가장 가까운 곳에 위치한 하수관거를 추출하고 전체 하수관로 데이터에 지반함몰 발생 유무를 표기하였다. 또한, 정확한 분석을 위해서 전체 데이터셋 중 결측치(Missing Value)와 오표기된 값을 제거한 후 분류 모델에 활용하였다. 정제된 데이터셋은 총 187,530개이며, 이 중 2,115개의 지반함몰 발생된 하수관로 데이터로 구성되어 있다.

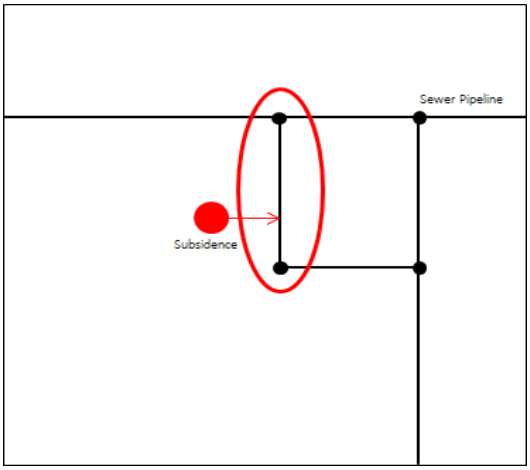


Fig. 5. Combination of sewer pipi and ground subsidence data using ArcGIS

Table 2. Summary of sewer pipe features

Category	Unit	Average	Standard error	Median	Min	Max
Age	Year	27.08	0.03	28.00	5.00	82.00
Slope	Degree	0.022	0.0002	0.007	0	16.31
Depth	Meter	1.00	0.001	0.85	0.01	147.41
Diameter	Millimeter	548.23	0.36	450.00	100.00	2,800.00
Length	Meter	27.72	0.04	25.11	0.02	1140.00

본 연구에서는 지반함몰 위험도에 영향을 미치는 인자들의 중요도를 확인하기 위해서 기존 연구결과를 활용하여 하수관로 속성정보 중 지반함몰 영향인자로 매설년수, 경사도, 매설심도, 관 직경, 관 길이로 선정하였으며 각 항목들의 기술통계량은 Table 2와 같다(Kim et al., 2017).

3.2 데이터 SET 전처리

정확한 분석결과를 취득하기 위해서는 원자료(raw data)를 데이터 분석의 목적과 방법에 맞는 형태로 처리하고 불필요한 정보를 분리·제거하고 가공하기 위한 전처리 작업이 필수이다. 본 연구에서는 하수관로 데이터와 지반함몰 데이터셋을 기계학습 알고리즘에 적용하기 위해서 Scaling과 같은 전처리 작업을 수행하였다. Scaling 기법은 데이터의 분포가 각기 다를 경우 일정 범위로 변경해주는 전처리 작업이며, 본 연구에서는 표준화(Standardization)을 통하여 단위가 각기 다른 데이터를 적절한 범위로 조정해주었다. 분석에 활용된 데이터셋은 하수관 특성 데이터와 지반함몰 발생유무를 결합한 데이터이다.

4. 기계학습 알고리즘 적용 및 결과

4.1 기계학습 적용

본 연구에서는 Python과 Scikit-learn 라이브러리를 활용하여 KNN, RF, LightGBM 알고리즘 중 하수관 특성에 따른 지반함몰 발생 분류에 적합한 모델을 선정하고자, 구축된 데이터셋을 통해 알고리즘에 학습한 뒤 평가지표를 비교하였다. 알고리즘 학습에 필요한 학습(Train) 데이터와 학습된 모델을 평가하기 위해서는 검증(Test) 데이터로 데이터셋을 분리하여 알고리즘에 적용해야 하며 본 연구에서는 학습-검증 비율을 80:20으로 설정하였다.

기계학습 모델 성능의 향상을 위해서는 문제 해결에 적절한 Hyper-parameter를 설정해야 한다. 본 연구에서는 다양한 Hyper-parameter의 조건에 대해 시행착오를 통해 최적의 성능을 탐색하는 GridSearchCV 알고리즘을 활용하여 주요 Hyper-parameter의 최적 성능 값을 선정하였다. GridSearchCV에 적용한 KNN 모델의 주요 Hyper-parameter 범위는 neighbor (1-3)이며, RF 모델은 estimators(100), max depth(6-12), min samples leaf(8-18), min samples split(8-20)으로, LightGBM 모델은 estimators(5-400), max depth(3-10), Learning rate(0.1), max iter(100)으로 설정하였다. 또한, 위의 조건을 GirdSearchCV 알고리즘에 적용하여 도출된 최적 성능 값을 Table 3에 나타냈다.

4.2 기계학습 결과

선정된 매개변수를 통해 전처리가 수행된 데이터셋을 적용한 결과를 Table 4에 제시하였다.

분석 결과 KNN 모델의 정확도는 높게 나타났으나, 정밀도와 재현율의 수치가 상대적으로 낮으며 모델의 성능을 나타내는

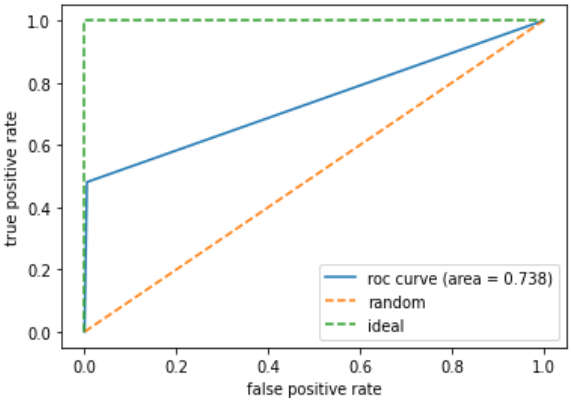
Table 3. Summary of hyper parameters in the model

Model	Hyper parameter
KNN	n neighbors(1)
RF	n_estimators(100), Max depth(12), Min samples split(16), Min samples leaf(8)
LightGBM	n_estimators(400), Max depth(10), Learning rate(0.1), Max iter(100)

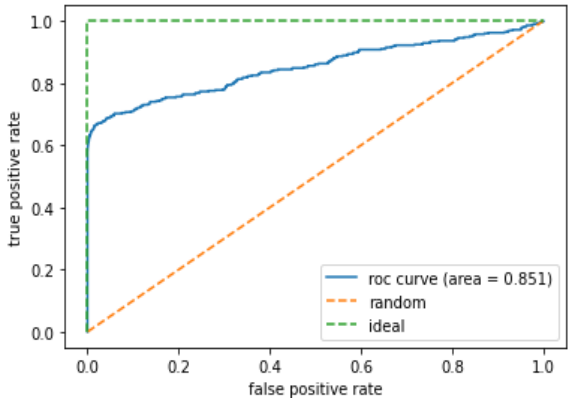
Table 4. Summary of hyper parameters in the model

Model	Accuracy	Precision	Recall	F1-Score	AUC
KNN	0.990	0.473	0.481	0.477	0.738
RF	0.995	130	0.579	0.733	0.851
Light GBM	0.993	0.666	0.712	0.688	0.852

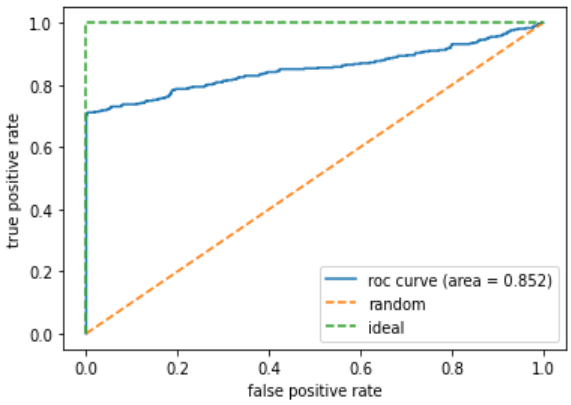
지표인 AUC의 값이 0.8 이하로 나타나 하수관 특성에 따른 지반함몰 발생 분류 모델에 적합하지 않은 것으로 판단된다. 다른 모델인 RF 모델은 실제 지반함몰 발생 데이터 중 바르게 지반함몰로 분류한 비율이 60%이며, AUC의 값이 0.851로 0.8보다 높아 지반함몰 발생 분류 모델에 적합한 것으로 판단된다. 마지막으로 LightGBM은 RF와 유사한 평가지표가 도출되었으며, AUC는 RF에 비하여 높은 것으로 나타났다. 따라서 하수관 특성에 따른 지반함몰 발생 분류 문제는 RF와 LightGBM 모델이 적합하며, 특히 RF 모델이 가장 우수한 성능을 나타내는 것으로 판단된다.



(a) KNN



(b) RF



(c) LightGBM

Fig. 6. ROC Curve of model

기계학습 모델은 과적합(Over-Fitting)에 대한 위험이 존재하기 때문에, 이에 대한 검증이 반드시 이루어져야 한다. 본 연구에서 선정된 RF 모델의 과적합 여부를 판단하기 위하여 Stratified K-Fold Cross Validation를 10회 반복하여 실시한 결과, 평균 0.996으로 나타나 과적합을 회피한 것으로 판단된다. Fig. 6은 KNN, RF, LightGBM의 ROC Curve를 나타낸 그래프이다.

지반함몰 발생 예측 결과, F1 Score가 가장 우수한 RF 모델의 결과를 분석해 보면 지반함몰 예측에 활용된 5개 인자들의 활용도, 즉 중요도는 Fig. 7과 같다. RF 모델은 경사도와 매설 년수를 집중적으로 활용하여 지반함몰 발생 분류를 예측하였고, 지반함몰에 영향을 미치는 하수관 특성의 중요도는 경사도, 매설 년수, 관의 길이, 평균심도, 관의 직경 순으로 나타냈다. RF 모델에서 도출된 경사도, 매설 년수가 하수관로 주변 지반함몰에 가장 큰 영향을 미치는 것으로 사료된다.

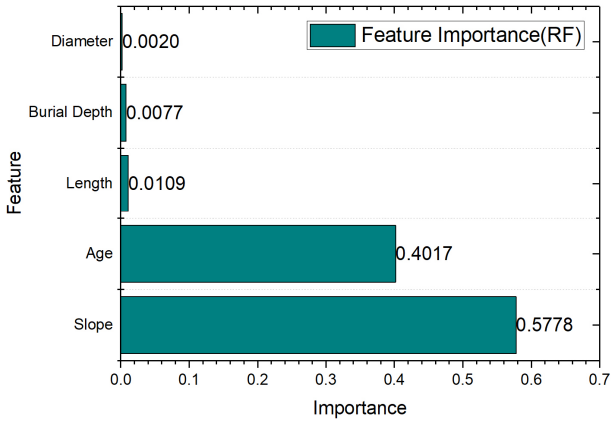


Fig. 7. Feature importance of RF

5. 결 론

본 연구는 OO시 하수관 특성과 지반함몰 발생 데이터를 활용하여 데이터셋을 구축하였으며, 구축된 데이터셋을 기계학습의 분류 알고리즘인 KNN, RF, LightGBM에 적용하여 하수관 특성에 따른 지반함몰 발생 분류 예측에 적합한 모델을 선정하고자 하였다. 또한, 선정된 모델을 통하여 분류 문제 해결에 활용된 주요한 입력 데이터의 중요도를 도출하였다. 본 연구의 요약과 결론은 다음과 같다.

(1) 하수관 특성에 따른 지반함몰 발생 예측 모델의 선정을 위하여 ArcGIS를 통해 데이터셋을 구축한 결과, 전체 하수관 데이터 중 지반함몰에 영향을 미친 것으로 가정한 데이터의 비율은 약 1.13%로 나타났다.

(2) 구축된 데이터셋의 전처리 과정을 수행한 후 최적화된 기계학습 알고리즘(KNN, RF, LightGBM)에 적용한 결과 상대적으로 RF와 LightGBM 모델의 평가 지표가 우수한 것으로 나타났으며, 모델의 성능 지표인 AUC가 0.8 이상으로 나타나 지반함몰 영향인자 분석을 위한 적절한 모델로 선정하였다.

(3) 분류 예측의 적절 모델로 선정된 RF모델을 통해 지반함몰에 주요한 영향을 미치는 하수관의 특성을 도출한 결과 경사도와 매설 년수를 중요한 영향인자로 판단하여 문제를 해결한 것으로 나타났다.

본 연구를 통해 제안된 기계학습 기법을 통한 하수관로 주변 지반함몰 예측 모델은 다양하고 복잡한 원인으로 발생하는 지반함몰 예측 모델의 기초 자료로 활용될 것으로 기대되며, 추가 연구를 통하여 다른 지역에 적용성 검증이 필요할 것으로 판단된다. 이와 같은 연구를 통해 하수관 주변 지반함몰 예방을 위한 관로 보수공사 시 우선순위 선정에 참고자료로 활용이 가능할 것으로 사료된다.

감사의 글

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

References

1. Breiman, L. (2001), Random Forests. Machine Learning, 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. Lee, H. M., Jeon, G. S. and Jang, J. A. (2020), Predicting of the severity of car traffic accidents on a highway using light gradient boosting model, The Journal of the Korea Institute of Electronic Communication Sciences Vol. 15, No. 6, pp. 1123~1130 (In Korean).
4. 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).
5. Kim, J. Y., Kang, J. M., Choi, C. H. and Park, D. H. (2017), Correlation analysis of sewer integrity and ground subsidence, Journal of the Korean Geo-Environmental Society, Vol. 18, No. 6, pp. 31~37 (In Korean).
6. 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.
7. Seoul Institute. (2016), The Road Subsidence Conditions and Safety Improvement Plans in Seoul (In Korean).
8. Seoul Metropolitan City. (2017), Road management technology white paper (In Korean).
9. Tom Fawcett (2005), An introduction to ROC analysis, Patter Recognition Letters, Edited by Francesco Tortorella, Vol. 27 Issue 8, pp. 861~874.