The Journal of the Convergence on Culture Technology (JCCT)
Vol. 9, No. 6, pp.145-150, November 30, 2023. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2023.9.6.145

JCCT 2023-11-18

머신러닝 기반 2호선 출퇴근 시간대 지하철 역사 내 혼잡도 예측

Subway Line 2 Congestion Prediction During Rush Hour Based on Machine Learning

장진영*, 김채원**, 박민서***

Jinyoung Jang*, Chaewon Kim*, Minseo Park**

요 약 지하철은 사람들이 일상적으로 이용하는 대중교통으로 자리잡고 있다. 특히 2호선은 지하철 승객이 하루동안 가장 많이 이용하는 역들이 포함되어 있는 호선으로 출퇴근 시간대에는 높은 혼잡도로 인해 압사사고의 위험성이 높아지고 있으며, 이는 지하철을 이용하는 사람들의 안전성과 쾌적함을 저하시킨다. 따라서 지하철 역사 내 혼잡도 예측을 바탕으로 높은 혼잡도로 인해 발생하는 문제를 대비할 필요가 있다. 이를 위해 본 연구에서는 출퇴근 시간대 혼잡 여부를 판별하는 머신러닝 분류 모델을 제안한다. 선행연구를 통해 지하철 혼잡도에 영향을 주는 변수를 파악하고, 공공데이터포털에서 출퇴근 시간대의 2호선 지하철 혼잡도 데이터셋을 수집하여 머신러닝을 기반하여 2호선 지하철 역사 내 혼잡 여부를 예측한다. 본 연구에서 제안하는 출퇴근 시간대 2호선 역사 내 혼잡도 예측 모델은 지하철 이용객의 안전과 만족도를 향상시키기 위한 지하철 운영 계획 수립에 활용될 수 있을 것으로 기대된다.

주요어 : 지하철, 2호선, 출퇴근 시간대, 혼잡도, 머신러닝, 지도학습

Abstract The subway is a public transportation that many people use every day. Line 2 especially has the most crowded stations during the day. However, the risk of crush accidents is increasing due to high congestion during rush hour and this reduces the safety and comfort of passengers. Subway congestion prediction is helpful to forestall problems caused by high congestion. Therefore, this study proposes machine learning classification models that predict subway congestion during commuting time. To predict congestion in Line 2 based in machine learning, we investigate variables that affect subway congestion through previous research and collect a dataset of subway congestion on Line 2 during rush hour from PUBLIC DATA PORTAL. The proposed model is expected to establish the subway operation plane to make passengers safe and satisfied.

Key words : Subway, Line 2, Rush Hours, Congestion, Machine Learning, Supervised Learning

Ⅰ. 서 론

이태원 참사 사건 이후 국내에서도 압사 사고에 대한 경각심이 높아지고 있다. 특히, 출퇴근 시간대를 중심으로 지하철 이용시 호흡곤란을 호소하는 승객이 지속적으로 발생하고 있다. 2022년 기준 일 평균 대중교통 이용

*준회원, 서울여자대학교 데이터사이언스학과 학부생(제1저자)
**준회원, 서울여자대학교 데이터사이언스학과 학부생(참여저자)
***정회원, 서울여자대학교 데이터사이언스학과 조교수 (교신저자)
접수일: 2023년 10월 4일, 수정완료일: 2023년 10월 20일
게재확정일: 2023년 11월 5일

Received: October 4, 2023 / Revised: October 20, 2023
Accepted: November 5, 2023
***Corresponding Author: mpark@swu.ac.kr
Dept. of Data Science, Seoul Women's Univ, Korea

건수는 약 1,025만 건이며, 지하철은 516만 건으로[1] 대중교통을 이용하는 서울시민 중 절반 이상이 지하철을 이용한다. 특히, 본 연구에서 분석 대상으로 선정한 2호선은 연평균 이용현황이 전체 지하철 이용 수요의 23.5%를 차지하고 있으며, 일별 최다 이용 지점(강남역, 잠실역, 홍대입구역 등)을 포함하고 있다[1].

역사 내 혼잡도는 철도 이용객의 만족도와 철도 운영에 큰 영향을 주는 요인이다. 혼잡도가 증가할 경우 이용객의 불쾌감은 상승할 수 있으며, 인적 사고로 인해 열차가 정차하는 등 철도 운영에 영향을 줄 수 있다[2]. 이에 따라 역사 내의 혼잡도를 완화하기 위한 연구가 활발하게 이루어지고 있다. 하지만, 주로 눈으로 관측하는 방식을 통해 지하철 혼잡도를 측정하였기에 구체적인 지하철 혼잡도 측정 지표가 부족한 실정이다. 2호선의 경우 예외적으로 실시간 하중 시스템을 통해 지하철 칸 별 혼잡도 정보를 제공하고 있지만, 이 역시 측정 무게와 실제 인원수 간의 오차가 존재하는 등 역사 내 혼잡도를 측정하는 데에는 한계를 지닌다.

따라서 본 연구에서는 머신러닝 기법을 적용하여 출퇴근 시간대 지하철 역사의 혼잡도를 예측하는 방법을 제안한다. 기존의 지하철 혼잡도 측정 방식 대신 환승역 개수, 환승객 수, 출입구 수 등 지하철 역사와 관련된 요소들을 기반하여 역사 내 혼잡도를 측정하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 지하철 혼잡도와 관련된 기존 연구를 서술하고, 3장에서는 분류에 효과적인 머신러닝 알고리즘을 소개한다. 4장에서는 본 연구에서 채택한 역사 내 혼잡도 예측 시스템과 연구 결과를 설명하고, 5장에서는 결론을 기술한다.

Ⅱ. 지하철 혼잡도 관련 선행연구

본 연구를 진행하기에 앞서 지하철 혼잡도에 영향을 미치는 요인을 파악하기 위해 선행연구를 조사하였다. 김동욱 외(2013)[3]와 김진수(2016)[4]에서는 혼잡도에 영향을 미치는 요인으로 승하차 인원수, 출구 수, 환승노선 수, 버스노선 수, 500m 이내 버스정류장 수, 500m 이내의 편의 시설 수가 활용되고 있다[3][4].

김진수[4]의 연구에서는 인천 지하철 1호선을 대상으로 다중회귀 모델을 이용하여 지하철 혼잡도 예측을 진행하였다. 승하차 인원 수, 여객 수입, SNS에 검색된 게시글 수 등을 독립변수로 사용하였고, SNS를 통해 얻은 유의미한 키워드를 기반으로 가중치를 적용하였다. 그러나 지하철이 환승역과 같은 중심역을 향하는 노선일 경우, 지하철 혼잡도는 증가하는 경향을 보였지만 이를 예측하는 모델의 성능은 대체적으로 낮아졌다. 이를 보완하기 위해 중심역과 각 방향에 따른 추가적인 분석이 필요하다[4].

김미례, 조인호[5]는 지하철 혼잡도 평준화를 위한 IoT(Internet of Things) 기술을 접목한 하중센서 측정 시스템을 제시하였다. 열차가 이전 역에서 정차하였을 때 하중 센서를 통해 승객 하중을 측정하고 이를 탑승 인원 수로 산정한다. 이를 기반으로 혼잡도를 계산하고, 무선 원거리 통신으로 다음역의 전광판에 혼잡도(매우 혼잡, 혼잡, 보통, 여유)가 표시되는 시스템이다. 혼잡도에 따라 게이트 수를 조절하여 승객이 몰리는 것을 방지하는 데는 유용하지만, 하중 센서로 무게를 측정하고 탑승 인원수를 계산하는 과정에서 여성과 남성의 무게 오차, 자전거 및 개인 짐과 같은 요인으로 인해 오차가 발생하는 문제가 있다[5].

백정현, 김치수 외[6]는 영상처리를 활용한 지하철 혼잡도 추정을 진행하였다. 지하철 내 카메라와 좌석 안의 압력센서를 통해 인원수 데이터를 수집하고, 이를 토대로 혼잡도 관련 정보를 제공한다. 영상처리를 통한 얼굴인식 방법과 카메라 앵글을 각 영역으로 나눈 뒤 머리 형태를 인식한다. 그러나, 얼굴인식 기술의 경우 원거리에서는 많은 제약이 있으며, 무엇보다 데이터 수집 과정에서 발생하는 초상권 및 개인정보 침해 문제가 우려된다는 한계점이 존재한다[6].

기존 연구에서는 다중회귀 모델, 영상처리 등 다양한 방법론이 활용되고 있으며, 지하철 이용객 수 및 지하철 칸 별 하중 정보를 이용하여 지하철 혼잡도를 예측하고 있다. 그러나 시간대에 따른 혼잡도 차이를 고려해 역사 내 혼잡도를 예측하는 연구는 부족한 실정이다. 이에 본 연구는 많은 사람들이 붐비는 출퇴근 시간대를 중심으로 머신러닝 기반 지하철 역사 내 혼잡도 예측 모델을 제안하고자 한다.

Ⅲ. 대표적인 분류 머신러닝 알고리즘

본 연구에서는 지하철 혼잡도를 바탕으로 2가지 범주(비혼잡/혼잡)로 데이터를 라벨링하였기에 분류 머신러닝 알고리즘을 적용하였다. 본 장에서는 분류 및 예측에 효과적인 머신러닝 알고리즘인 로지스틱 회귀, 의사결정나무, 랜덤 포레스트를 살펴보고자 한다[7].

1. 로지스틱 회귀

로지스틱 회귀는 라벨링한 데이터를 학습한 후 새로운 데이터가 두 개의 범주 중 어느 쪽에 속하는지 판별하는 알고리즘이다[7]. 새로운 데이터가 두 그룹 중 하나에 속할 확률을 계산하고 이를 토대로 데이터를 분류한다[8].

2. 의사결정나무

의사결정나무는 학습한 데이터를 토대로 규칙을 생성한 뒤 불순도가 낮아지도록 데이터를 분류하는 알고리즘이다[9]. 입력된 데이터의 변수(Feature)를 토대로 불순도를 감소시키는 방향으로 데이터를 분류하기 때문에 어떤 변수가 분류 기준으로 중요하게 작용했는지 확인할 수 있다[9][10].

3. 랜덤 포레스트

랜덤 포레스트는 하나의 결정나무가 아닌 여러 개의 나무를 생성한 후 데이터를 분류하는 알고리즘이다[11]. 데이터를 무작위로 선정하여 학습 데이터의 부분 집합으로 여러 개의 나무를 생성하기 때문에 하나의 나무를 기반으로 하는 의사결정나무와 달리 학습데이터가 편향되는 것을 방지할 수 있다[12].

Ⅳ. 2호선 지하철 혼잡도 예측

본 연구는 선행연구를 기반으로 수집한 변수들을 기반으로 출퇴근 시간대의 역사 내 혼잡도 예측 모델을 생성하였다. 생성한 모델의 결과를 해석하여 출퇴근 시간대의 혼잡도를 예측하는 데에 효과적인 지표를 탐색하고자 한다. 전체적인 프로세스는 그림 1과 같다.

1. 데이터 수집

본 연구에서는 공공데이터포털에서 제공하는 서울교통공사 혼잡도 정보 2019년과 2022년 데이터셋을 활용하였다. 2020년과 2021년의 경우 코로나19로 지하철 이용객이 예외적으로 저조하여 해당년도는 제외하였다. 기존 연구를 통해 채택한 변수인 지하철 환승 노선 수, 500m이내 버스 정류장 수, 지하철 출입구 수, 건물 수, 승하차 인원 수에 대한 데이터를 수집하였다. 또한, 본 연구는 역사 내 혼잡도 예측을 목표로 하기 때문에 대합실 면적, 승강장 면적 등과 같은 추가적인 데이터를 수집하였다. 최종적으로 머신러닝 학습을 위해 사용한 데이터셋에 포함된 변수들은 표 1과 같다.

2. 데이터 전처리

우선 문자형 변수(Object)를 수치형 변수(Numeric)로 변환해주는 과정이 필요하다[13]. 수집한 데이터셋에서 문자형 변수인 출발역의 경우, 0과 1을 사용하여 전처리하였다. 예를 들어 강남역을 나타내는 컬럼이면 강남역은 1, 나머지 역은 0으로 표현하였다. 상하구분도 내선을 1, 외선을 0으로 변경하여 구분하였다. 또한 각 독립변수의 범위가 다를 경우, 종속변수에 미치는 영향의 정도가 달라지기 때문에[14] 범위를 맞춰주는 정규화 과정을 통해 범주형 데이터를 제외한 모든 수치형 변수의 평균을 0, 분산을 1로 범위를 통일하였다[14].

추가적으로 종속변수에 해당하는 지하철 혼잡도를 비혼잡, 혼잡으로 나누어 범주형 변수(Categorical)로 생성하였다. 국토교통부의 혼잡도 기준에 따라'주의'에 해당하는 130%을 기준으로 구분하였다. 보통은 여유롭게 이동 가능한 상태, 주의는 이동 시 부딪힘이 있는 상태, 혼잡은 열차 내 이동이 불가능한 상태를 의미한다[15]. 이동 시 부딪힘이 있는 상태는 승객의 안전에 영향을 미칠 수 있다고 판단하였고, 이에 기준을 주의로 설정하였다. 따라서 본 연구에서는 130% 미만이면 비혼잡으로, 이상이면 혼잡으로 이진 분류로 구성하였다.

3. 모델링

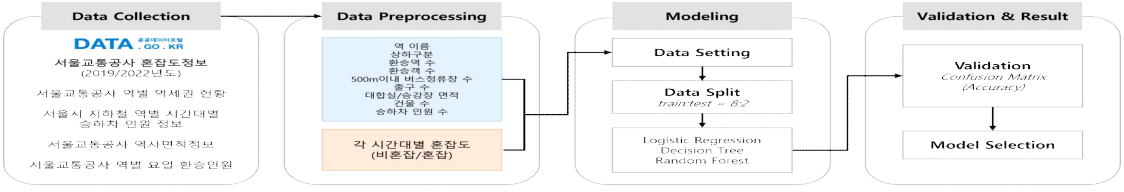

그림 1. 지하철 혼잡도 예측을 위한 순서도

Figure 1. Flow Chart of Subway Congestion Prediction

표 1. 변수의 정의
Table 1. Definition of Variables

변수명		정의	유형	출처
독립변수	역 이름	출발역 이름	Categorical	서울교통공사 혼잡도 정보(2019/2022년도)
	상하구분	지하철의 운행 방향(상선 또는 하선)	Categorical	
	환승역 수	해당 역의 환승 노선 수	Numeric	직접 수집
	환승객 수	해당 역의 평일 일평균 환승객 수	Numeric	서울교통공사 역별요일 환승인원
	버스정류장 수	반경500m이내 버스정류장 수	Numeric	직접 수집
	출입구 수	해당 역의 출입구 수	Numeric	직접 수집
	대합실 면적	해당 역의 대합실 면적	Numeric	서울교통공사 역사면적정보
	승강장 면적	해당 역의 승강장 면적	Numeric	
	건물 수	해당 역의 역세권 건물 수	Numeric	서울교통공사 역별역세권 현황
	승차인원 수	해당 역의 시간대별 승차인원 수	Numeric	서울시 지하철 역별 시간대별 승하차 인원 정보
	하차인원 수	해당 역의 시간대별 하차인원 수	Numeric	
종속변수	시간대별 혼잡도(비혼잡/혼잡)	특정 시간대의 혼잡도 (130% 미만은 비혼잡, 130% 이상은 혼잡)	Categorical	서울교통공사 혼잡도 정보(2019/2022년도)

전처리한 데이터셋을 8:2의 비율로 학습 데이터와 테스트 데이터로 나누어 분류에 효과적인 알고리즘인 로지스틱 회귀, 의사결정나무, 랜덤 포레스트 알고리즘으로 학습하였다. 각각의 출퇴근 시간대를 기반으로 학습하였고 영향력이 큰 변수를 확인하였다. 또한 시간대별 최적의 모델을 탐색하기 위해 하이퍼파라미터 값을 조정하였다. 하이퍼파라미터는 머신러닝 알고리즘의 성능을 향상시키기 위해 튜닝을 통해서 결정되는 변수를 의미하며, 값에 따라 모델의 성능이 달라진다[16]. 의사결정나무와 랜덤 포레스트의 max_depth는 결정나무의 최대 깊이를 설정하는 하이퍼파라미터를 의미한다. 그러나 max_depth 값이 지나치게 클수록 학습 데이터를 지나치게 학습될 가능성도 있다[17]. 본 연구에서는 이러한 과적합을 방지하기 위해 결정나무의 깊이를 3으로 설정하였다. 랜덤 포레스트의 n_estimators은 생성할 결정나무의 개수를 결정하는 하이퍼파라미터를 의미한다. 결정나무의 개수가 충분해야 성능이 높아지지만 모델의 복잡성을 증가시킬 수 있기에 적절한 값을 설정하는 것이 중요하다[18]. 본 연구에서는 500으로 설정하였다[19].

4. 모델 성능 평가 및 결과

전처리한 데이터셋을 가장 효과적인 지하철 혼잡도 예측 알고리즘을 선정하기 위해 정확도(Accuracy)를 기준으로 세 가지 모델의 성능을 비교하였다. 표 2는 생성한 출퇴근 시간대 모델의 성능을 나타낸다.

정확도는 수식(1)과 같다. 본 연구에서 정확도는 지하철 혼잡도의 실제 값과 예측한 값이 정확히 일치한 경우를 의미한다.

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

오후 6시 30분을 제외한 나머지 시간대에서는 의사결정나무를 예측 모델로 선정하였으며, 오후 6시 30분 모델로서 랜덤 포레스트를 채택하였다. 이는 학습에 사용한 데이터의 크기가 작아 하나의 의사결정나무로도 데이터를 분류하는 데에 충분함을 알려준다[10]. 오전 8시와 오전 9시의 경우, 의사결정나무와 랜덤 포레스트는 테스트 데이터에서 0.91로 동일한 성능을 보이지만 변수의 개수가 적은 데이터에 효과적이며, 모델 학습 결과를 가시적으로 확인할 수 있는[20] 의사결정 나무를 제안한다. 또한, 오전 8시 30분 모델에서 랜덤 포레스트가 의사결정나무에 비해 학습 데이터의 정확도가 더 높으나 테스트 데이터의 정확도가 낮아지는 것을 보아 학습 데이터를 지나치게 학습한 것을 확인할 수 있다. 이에 따라 8시 30분 모델로서 랜덤 포레스트가 아닌 의사결정나무를 또한 제안한다.

각 모델의 변수 중요도를 출력한 결과 출근 시간대인 오전 8시 모델의 경우 상하구분, 500m 이내 버스정류장 수 순으로 나타났으며, 오전 8시 30분 모델의 경우 승차인원 수, 건물 수, 하차 인원 수, 상하구분 순으로 나타났다. 오전 9시 모델의 경우 하차 인원 수, 상하구분 순으로 나타났다. 출근시간대의 경우 공통적으로 상하구분이 혼잡도를 예측하는 데에 있어 중요한 지표로 나타났음을

알 수 있다. 퇴근 시간대인 오후 6시 모델의 경우 승강장 면적, 상하구분 순으로 변수 중요도가 출력되었다. 오후 6시 30분 모델의 경우 랜덤 포레스트가 혼잡도 예측을 위한 모델로 선정되었기에 각 변수의 중요도가 다른 의사결정나무에 비해 고르게 출력되었지만, 하차 인원 수, 건물 수, 출구 수, 상하구분이 다른 변수에 비해 중요도가 높음을 알 수 있다.

원 수, 건물 수, 출구 수, 상하구분이 비교적 높게 나타났다. 상하구분은 출퇴근 시간대 모두에서 높은 중요도를 보였으며, 승강장 면적은 퇴근 시간대에서만 중요한 변수로 나타났다.

	오전 8:00 (train)	오전 8:00 (test)	오전 8:30 (train)	오전 8:30 (test)	오전 9:00 (train)	오전 9:00 (test)	오후 6:00 (train)	오후 6:00 (test)	오후 6:30 (train)	오후 6:30 (test)
Logistic Regression	0.94	0.8	0.94	0.77	0.94	0.83	0.92	0.86	0.97	0.86
Decision Tree	0.95	0.91	0.98	0.94	0.98	0.91	0.95	0.88	1.0	0.94
Random Forest	0.97	0.91	0.99	0.86	0.97	0.91	0.97	0.83	0.98	0.97

V. 결 론

본 연구에서는 서울교통공사에서 제공하는 지하철 혼잡도 정보를 활용해 출퇴근 시간대 역사 내 혼잡도 예측 머신러닝 모델을 제안하였다. 데이터 수집과 전처리 과정을 통해 역 이름, 상하구분, 환승역 수 등을 포함한 총 11개의 변수를 독립변수로 채택하였으며, 비혼잡과 혼잡으로 나뉘는 시간대별 혼잡도를 종속변수로 사용하였다.

혼잡도(비혼잡/혼잡)를 예측하기 위해 대표적인 분류 알고리즘인 로지스틱 회귀, 의사결정나무, 랜덤 포레스트를 활용하여 모델링을 진행하였다. 분석 결과 오전 8시, 오전 8시 30분, 오전 9시, 오후 6시에서는 정확도가 각각 0.91, 0.94, 0.91, 0.88로 의사결정나무에서 가장 높은 성능을 보였다. 오후 6시 30분에서는 랜덤 포레스트에서 0.97의 정확도를 보였다. 이에 각 출퇴근 시간대 최적의 모델로서 오후 6시 30분은 랜덤 포레스트로, 이외의 시간대는 의사결정나무로 채택하였다. 또한 출퇴근 시간대 혼잡도에 영향을 주는 요인을 확인하고 비교하기 위해 각 변수의 중요도를 살펴보았다. 출근 시간대인 오전 8시에서는 상하구분이 가장 높게 나타났으며 다음으로는 500m 이내 버스 정류장 수의 중요도가 높게 나타났다. 오전 8시 30분에서는 승차 인원수, 건물 수, 하차 인원 수, 상하구분 순으로 나타났으며, 오전 9시에서는 하차 인원 수, 상하구분 순으로 나타났다. 퇴근 시간대인 오후 6시의 경우는 승강장 면적, 상하구분 순으로 높게 나타났으며, 오후 6시 30분에서는 하차 인

본 연구는 혼잡도 예측을 위한 새로운 지표를 제공할 수 있다는 점에서 의의가 있다. 해당 연구에서 제안하는 지하철 혼잡도 측정 지표는 혼잡도 변동량에 따라 열차 배차간격을 길게 하거나 짧게 조정하여 승객이 몰리는 것을 방지하여 승객의 안전을 보장하고, 승객의 열차 이용 만족도를 향상하는 데에 활용이 가능할 것으로 기대된다. 그러나, 본 연구에는 몇 가지 한계점이 존재한다. 해당 연구에서는 2호선을 대상으로 진행했기에 연구 결과를 다른 호선에 일반화하기엔 무리가 있다. 해당 연구 결과를 전체 노선에 일반화하기보단 각 노선별 특징을 반영하여 추가적인 분석을 진행할 필요가 있다. 또한 해당 연구는 2019년과 2022년, 출퇴근 시간대만을 가진 샘플 약 200개만을 사용하였기에 더 많은 샘플을 확보하여 연구 결과를 보완할 필요가 있다.

References

[1] Futuristic Advanced Transportation Division of Seoul City Transportation Office,『Seoul Transportation in 2022』, 2023.
[2] S.-H. Lee, C.-K. Cheon, B.-D. Jung, B.-Y. Yu, andE.-J. Kim.“Study on Methodology for Effect Evaluation of Information Offering to Rail passengers”, *The Journal of The Korea Institute of Intelligent Transport Systems*, Vol. 14, No. 3, pp. 50-62, 2015. DOI:10.12815/kits.2015.14.3.050
[3] Dong-Wook Kim, Un-yong Kim, Jun-Won Lee,“Application of Multiple Regression Analysis to Improve Congestion in Subway Carriages and Design of Optimization Plan”, *Korean Institute of Industrial Engineers*, Vol. 2013, No. 11, pp. 1441-1448, 2013.
[4] Kim, Jin-Su. “Subway Congestion Prediction

and Recommendation System Using Big Data Analysis", *Journal of Digital Convergence*, Vol. 14, No. 11, pp. 289 - 295, 2016. DOI:10.14400/JDC.2016.14.11.289

[5] Mi-Rye Kim, In-Ho Cho."Design of Congestion Standardization System Based on IoT", *Journal of the Korea Academia-Industrial Cooperation Society*, Vol. 17, No. 5, pp. 74 - 79, 2016, DOI:10.5762/KAIS.2016.17.5.74

[6] Jung-Hyun Back, Chi-Su Kim, Jun-Ha Park, Gun-Hee Ye, Dong-Soo Jang, Wook-Hyun Ha, and Dong-Kweon Hong. "Estimating subway congestion using image processing", *Korean Institute of Information Scientists and Engineers*, Vol. 2017, No. 12, pp. 533-535, 2017.

[7] Nasteski, V. "An overview of the supervised machine learning methods", *Horizons,* Vol. b, No. 4, pp. 51-62, 2017. DOI:10.20544/HORIZONS.B.04.1.17.P05

[8] S.-B. Jin ,J.-W. Lee, "Study on Accident Prediction Models in Urban Railway Casualty Accidents Using Logistic Regression Analysis Model", Journal of the Korean Society for Railway, Vol. 20, No. 4, pp. 482-490, 2017. DOI: 10.7782/JKSR.2017.20.4.482

[9] Kazemitabar, J., Amini, A., Bloniarz, A., and Talwalkar, A. S. "Variable importance using decision trees", *Advances in neural information processing systems*, Vol. 30, 2017.

[10]J. Hong, S.-J. Jeon, "Prediction of Safety Grade of Bridges Using the Classification Models of Decision Tree and Random Forest" *KSCE Journal of Civil and Environmental Engineering Research*, Vol. 43, No. 3, pp. 397 - 411, 2023. DOI:10.12652/Ksce.2023.43.3.0397

[11]Biau, G.,Scornet, E."A random forest guided tour", *Test*, Vol. 25, pp. 197-227, 2016. DOI:10.1007/s11749-016-0481-7

[12]JungIn Seo,JeongHyeon Chang. "Predicting Reports of Theft in Businesses via Machine Learning", *International Journal of Advanced Culture Technology(IJACT),* Vol. 10, No. 4, pp. 499-510, 2022. DOI:10.17703/IJACT.2022.10.4.499

[13]Al-Shehari, Taher, Rakan A. Alsowail, "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques", *Entropy,* Vol. 23, No. 10, pp. 1258, 2021. DOI:10.3390/e23101258

[14]Peshawa J. Muhammad Ali, Rezhna H.Faraj; "Data Normalization and Standardization: A Technical Report, *Machine Learning Technical Reports*, Vol. 1, No. 1, pp 1-6, 2014. DOI:10.13140/RG.2.2.28948.04489

[15]Seoul Metro(2021),"The Seoul Transportation Corporation won the '10 Best Technology Awards for Subway Congestion Calculation Service' for Big Data Convergence", https://www.seoul.go.kr/news/news_report.do#view/350703

[16]Seokjin Im,"An Extended Function Point Model for Estimating the Implementing Cost of Machine Learning Applications."*The Journal of the Convergence on Culture Technology,* Vol. 9, No. 2, pp. 475 - 481, 2023. DOI:10.17703/JCCT.2023.9.2.475

[17]S. Jiang, H. Mao, Z. Ding and Y. Fu, "Deep Decision Tree Transfer Boosting", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 31, No. 2, pp. 383-395, 2020, DOI: 10.1109/TNNLS.2019.2901273

[18]Garibay, Alonso Palomino et al. "A Random Forest Approach for Authorship Profiling" , *Proceedings of CLEF*, 2015.

[19]Meng D, Xu Jun, Zhao J, "Analysis and prediction of hand, foot and mouth disease incidence in China using Random Forest and XGBoost", *Plos one*, Vol. 16, No. 12, 2021. DOI:10.1371/journal.pone.0261629

[20]Safavian, S. R., & Landgrebe, D. "A survey of decision tree classifier methodology", *IEEE transactions on systems, man, and cybernetics,* Vol. 21, No. 3, pp. 660-674, 1991. DOI:10.1109/21.97458

※ 이 논문은 서울여자대학교 학술연구비의 지원에 의한 것임 (2023-0226).