https://doi.org/10.7840/kics.2025.50.6.847

지리 가중 회귀와 머신러닝을 활용한 데이터 센터 최적 입지 선정

이 우 진◆, 김 민 영*, 윤 선 영*, 김 수 현°

Optimal Data Center Location Selection Using Geographically Weighted Regression and Machine Learning

Woojin Lee◆, Minyoung Kim*, Sunyoung Yoon*, Suhyeon Kim°

요 약

데이터 센터는 디지털 경제의 핵심 인프라로, 급증하는 데이터 처리 수요와 함께 전력 소비 및 환경적 부담이 증가하고 있다. 특히, 한국에서는 데이터 센터가 수도권에 과도하게 집중되어 전력 수급 불균형과 지역 불균형을 심화시키는 주요 원인으로 작용하고 있다. 이러한 문제를 해결하기 위해 본 연구는 머신러닝과 지리 가중 회귀를 활용하여 데이터 센터 최적 입지를 선정하는 방법론을 제안한다. 먼저, 랜덤 포레스트를 통해 데이터 센터 입지에 영향을 미치는 주요 변수들의 중요도를 도출하였으며, 자연재해 위험(침수 위험, 지진 위험)과 인프라 조건(인구 밀도, 전력 공급 안정성)이 핵심 요인으로 확인되었다. 이후 지리 가중 회귀를 활용하여 지역적 특성을 반영한 회귀 계수를 산출하였으며, 이를 바탕으로 지역별 입지 적합성을 평가하였다. 분석 결과, 천안, 김해, 대구 등 일부 지방 지역이 데이터 센터 입지에 적합한 것으로 나타났다. 이들 지역은 자연재해 위험이 낮고 재생 가능 에너지 접근성과 인프라 조건이 우수하여 데이터 센터 운영의 안정성과 지속 가능성을 동시에 충족할 가능성을 보였다. 본 연구는 데이터 센터 입지 선정 과정에서 자연재해, 인프라 조건, 사회적 요인 간의 상호작용을 통합적으로 고려한 새로운 프레임워크를 제시하며, 데이터 기반 의사결정의 정교함을 높였다. 나아가 본 연구는 데이터 센터 뿐만 아니라 기타 주요 기반 시설의 입지 선정에도 활용될 수 있는 가능성을 제시하며, 지역 균형 발전과 지속 가능한 데이터 센터 운영을 위한 실질적 정책적 시사점을 제공한다.

키워드 : 데이터 센터 입지 선정, 랜덤 포레스트, 지리 가중 회귀, 입지 적합성 평가

Key Words : Data Center Location Selection, Random Forest, Geographically Weighted Regression, Location Suitability Assessment

ABSTRACT

Data centers serve as critical infrastructure in the digital economy, supporting the growing demand for data processing while contributing significantly to energy consumption and environmental impact. In South Korea,

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2023-00245529). Also, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2023-00242528) and by the IITP (Institute for Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (No. IITP-2025-RS-2024-00437756).
◆ First Author : Kyungpook National University, Graduate School of Data Science, leewoojin708@knu.ac.kr, 학생회원
° Corresponding Author : Kyungpook National University, Graduate School of Data Science, suhyeonkim@knu.ac.kr, 정회원
* Kyungpook National University, Graduate School of Data Science, tt08@naver.com; tjsdud6900@naver.com
논문번호 : 202501-005-A-RU, Received December 31, 2024; Revised February 1, 2025; Accepted March 19, 2025

the excessive concentration of data centers in metropolitan areas has exacerbated issues such as power supply imbalances and regional disparities. To address these challenges, this study proposes a methodological framework for optimal data center location selection by integrating machine learning and geographically weighted regression (GWR). Random Forest was employed to identify key factors influencing site suitability, revealing that natural disaster risks (e.g., flood and earthquake risks) and infrastructure conditions (e.g., population density and power supply stability) are critical determinants. GWR was subsequently utilized to estimate region-specific regression coefficients, incorporating local characteristics into the evaluation of location suitability. The analysis identified Cheonan, Gimhae, and Daegu as highly suitable locations, characterized by lower natural disaster risks, accessibility to renewable energy, and favorable infrastructure conditions, thereby ensuring operational stability and sustainability. This study advances the decision-making process by providing a comprehensive framework that considers the interaction between natural disaster risks, infrastructure conditions, and regional characteristics. Furthermore, the proposed methodology has potential applications in other critical infrastructure domains, offering practical insights for achieving regional balance and sustainable data center operations.

Ⅰ. 서 론

데이터 센터는 디지털 경제의 핵심 인프라로, 대규모 데이터 분석, 인공지능, 클라우드 컴퓨팅 등 현대 산업 전반에 걸쳐 중추적인 역할을 수행하고 있다[1,2]. 급증하는 데이터 처리 수요에 대응하며 기술 혁신의 기반을 제공하지만, 데이터 센터의 확산은 막대한 전력 소비와 환경적 부담을 초래하고 있다[3,4]. 현재 데이터 센터는 연간 약 200TWh의 전력을 소비하며, 이는 전 세계 전력 사용량의 약 1%를 차지한다[5]. 특히 냉각 시스템은 데이터 센터 전력 소비의 상당 부분을 차지하며, 전력 및 물 자원의 과도한 사용은 환경에 심각한 부정적 영향을 미친다[6,7]. 이러한 문제는 데이터 센터의 효율성과 지속 가능성을 개선하기 위한 입지 선정의 중요성을 부각시키고 있다.

한국에서는 데이터 센터가 수도권에 지나치게 집중되어 있어 전력 수급 안정성과 지역 균형 발전을 저해하고 있다[8]. 반면, 지방은 풍부한 재생 가능 에너지 자원과 낮은 토지 비용을 제공하여 데이터 센터 분산 배치의 잠재적 대안으로 부상하고 있다[9,10]. 수도권 중심의 데이터 센터 운영으로 발생하는 환경적, 경제적 문제를 완화하기 위해 적합성이 높은 타지역으로의 데이터 센터 분산 배치는 실효성 있는 해결책으로 제시되고 있다. 이는 단순히 물리적 공간의 확보를 넘어, 자연재해 위험, 인프라 조건, 지역 특성을 종합적으로 고려하는 정교한 접근을 요구한다[11,12].

기존 연구는 데이터 센터 입지 선정과 관련하여 다양한 분석과 모델을 제시해 왔다. 예를 들어, 자연재해 데이터를 기반으로 클라우드 데이터 센터의 입지 적합성을 평가하거나[5], 한국 공공 부문의 데이터 센터를 대상으로, 지속 가능성과 운영 효율성을 고려한 녹색 데이터 센터의 전략적 입지 설계를 탐구한 사례 연구를 제시해 왔다[13]. 그러나 기존 연구는 변수 간 상호작용을 충분히 반영하지 못하거나 특정 지역에 한정된 사례로 일반화 가능성에 제약이 있었다.

본 연구는 이러한 한계를 극복하기 위해 랜덤 포레스트(Random Forest)와 지리 가중 회귀(Geographically Weighted Regression, GWR)를 활용하여 데이터 센터 입지 적합성을 단계적으로 분석하는 새로운 접근법을 제안한다. 랜덤 포레스트를 통해 데이터 센터 입지에 영향을 미치는 주요 변수의 중요도를 도출하고, GWR을 통해 지역별 특성을 정량적으로 반영하여 최종적으로 입지 적합성 점수를 산출하는 통합 점수화 방식을 채택하였다. 이는 수도권 집중 문제를 완화하고 지방으로 데이터 센터를 분산 배치함으로써 전력 소비와 환경적 부담을 경감하며, 지역 균형 발전과 지속 가능한 운영을 실현할 수 있는 실질적인 해결책을 제시한다.

본 연구는 데이터 센터 입지 선정 과정에서 머신러닝 기반의 전역적인 변수 중요도 분석과 공간 회귀 모델의 지역적인 계수 정보를 결합하여, 지역별 입지 적합성을 정량적으로 평가하는 새로운 프레임워크를 제시한다. 본 연구의 결과는 향후 데이터 센터 입지 선정 및 관련 연구에서 데이터 센터 운영의 지속 가능성을 강화하기 위한 정책적 기초 정보, 분석 가이드라인 등으로 활용될 수 있다.

Ⅱ. 선행 연구

데이터 센터 입지 선정은 지속 가능한 데이터 센터 운영과 지역 균형 발전을 실현하기 위한 중요한 연구 주제로 다뤄져 왔다. 기존 연구들은 데이터 센터 입지 선정의 핵심 요소를 규명하고, 지역적 특성을 반영한 최적화 모델을 개발하며 관련 도구와 기법을 활용해왔다. 그러나 대부분의 연구는 개별 요인에 집중하거나 변수 간 상호작용 및 지역적 특성을 충분히 반영하지 못한 한계를 보였다.

선행 연구 중 하나로, 클라우드 데이터 센터의 입지 선정을 위해 재해 정보를 활용한 모델이 제안된 바 있다[5]. 이 연구는 자연재해 위험(e.g., 침수, 산사태, 해일 등)과 인적 재해(e.g., 가스 충전소, 주유소 등) 데이터를 활용하여 클라우드 데이터 센터의 안정성을 보장할 수 있는 최적 입지를 분석하였다. 그러나 이 연구는 변수 간 상호작용에 대한 분석이 미흡하다는 한계가 있다.

지속 가능한 데이터 센터 입지 선정을 위해 다기준 의사결정 분석을 활용하거나[14], 다기준 세트 커버링 모델이 제안되기도 하였다[15]. 전자는 경제적, 환경적, 사회적 요인을 활용하여 데이터 센터의 적합성을 분석하였으며, 비용 효율성과 환경적 영향을 최소화할 수 있는 대안을 제시하였다. 후자는 서비스 신청자와 데이터 센터 간의 거리, 재생 가능 에너지 접근성, 물리적 인프라와 같은 지속 가능성 요인을 주요 변수로 설정하여 분석을 진행하였다. 그러나 이러한 선행연구들은 변수 간 비선형적 관계를 설명하거나 지역적 특성을 정량적으로 반영하지 못한 점이 한계로 남아있다.

이와 같은 기존 연구들은 데이터 센터 입지 선정의 주요 요인을 분석하고 다양한 방법론을 적용했지만, 개별 변수에만 초점을 맞추거나 지역적 특성을 정량적으로 반영하지 못하는 한계가 있었다. 본 연구는 이러한 한계를 극복하기 위해 머신러닝 기반 변수 중요도 분석과 공간 회귀 모델의 분석 결과를 결합하여 전역적-지역적 요소를 반영한 입지 선정 분석 프레임워크를 제안한다. 랜덤 포레스트를 활용하여 데이터 센터 입지에 영향을 미치는 주요 변수의 상대적 중요도를 분석하고, GWR 모델을 통해 지역별 공간적 특성을 반영한 회귀계수를 산출함으로써 기존 연구에서 부족했던 변수 간 상호작용과 지역적 차이를 종합적으로 평가하였다. 특히, 본 연구는 자연재해 위험, 인프라 조건, 사회적 요인 간의 상호작용을 통합적으로 고려하여, 데이터 기반의 정밀한 입지 평가를 가능하게 하였다.

Ⅲ. 본 론

3.1 분석 데이터

본 연구는 데이터 센터 최적 입지 선정을 위해 기존 연구에서 중요한 요소로 다루어진 입지 선정 요인[5]들을 종합적으로 검토한 후, 다양한 기관들에서 데이터를 수집 및 통합하여 활용하였다. 분석에 사용된 데이터는 산림청, 행정안전부, 기상청, 국토교통부 등 총 8개 기관에서 제공하는 16가지 종류의 공공 데이터를 기반으로 구축되었다. 데이터는 재해 이력, 위험 지역, 지형 조건, 인프라 현황 등 총 16개의 변수를 포함하며, 이는 데이터 센터 입지 적합성을 다차원적으로 평가하기 위한 기초 자료로 활용되었다. 데이터 통합 과정에서는 행정구역 시군구 경계 데이터를 기준으로 모든 변수를 매핑하여 단일 데이터셋으로 구성하였다. 수집한 데이터의 세부 목록은 Table 1과 같다.

Table 1. Public Data Sources and Categories for Data Center Site Selection Analysis

Data Name	Organization
Landslide Information	Korea Forest Service
Nationwide Landslide Occurrence History (Last 5 Years)	Korea Forest Service
Standard Data on Nationwide Disaster Risk Zones	Ministry of Public Administration and Security
Statistical Yearbook: Regional Earthquake Occurrences	Ministry of Public Administration and Security
Earthquake Yearbook	Korea Meteorological Administration
Multipurpose Dam Operation Information	Korea Water Resources Corporation
Designation and Notification Status of Disaster-Risk Reservoirs	Ministry of Public Administration and Security
Nationwide Gas Station Registration Status	Ministry of Trade, Industry and Energy
Regional and Hourly Solar and Wind Power Generation	Korea Electric Power Exchange (KEPCO)
Wind Turbine Location Information	Korea Energy Agency
Disaster Risk Zones	Ministry of Land, Infrastructure and Transport
Land/Parcel Statistics	Ministry of Land, Infrastructure and Transport

Data Name	Organization
Total Population from National Geographic Information Institute	Ministry of Land, Infrastructure and Transport
Building Area from National Geographic Information Institute	Ministry of Land, Infrastructure and Transport
Administrative District Boundaries (City, County, District)	Ministry of Land, Infrastructure and Transport
Status of Public Institution Data Centers	Ministry of Public Administration and Security

본 연구에서 사용된 데이터 샘플의 총 개수는 약 230여 곳에 달하며, 이는 전국 시·군·구를 규모로 수집하였다(Table 2). 자료 수집 기간은 2020년부터 2023년까지이며, 이 기간 동안 공공데이터 포털, 지자체 통계청, 민간 보고서 등 다양한 출처에서 수집된 정보를 교차 검증함으로써 데이터의 신뢰도를 높이고자 하였다.

Table 2를 통해 확인한 바와 같이, 국내 데이터 센터는 수도권(서울, 경기, 인천)에 지나치게 집중되어 있으며, 전체 데이터 센터의 약 80%가 이 지역에 위치하고 있다. 반면, 경상북도, 전라북도, 제주도 등 일부 지역에는 데이터 센터가 전혀 분포하지 않은 것으로 나타

Table 2. Distribution of Data Centers by Region in South Korea (as of 2023)

Region	Count
Seoul	20
Busan	4
Gangwon-do	4
Gyeonggi-do	20
Gyeongsangbuk-do	0
Daegu	1
Gwangju	1
Daejeon	1
Sejong	1
Incheon	2
Jeollanam-do	1
Jeollabuk-do	0
Jeju Island	0
Chungcheongnam-do	1
Chungcheongnam-do	1
Ulsan	1
Gyeongsangnam-do	1

났다. 이러한 수도권 중심의 분포는 전력 수급 불균형 및 지역 간 발전 격차를 심화시킬 수 있는 구조적 문제로 지적된다.

본 연구는 이러한 편향된 분포 문제를 해결하기 위해 데이터 센터 입지 적합성을 평가할 수 있는 16개의 변수(Table 3)를 설정하였다. 변수 선정 과정에서 국내외 선행 연구를 기반으로 데이터 센터 입지 선정에 중요한 요인으로 반복적으로 제시된 변수를 우선적으로 고려하였으며, 데이터의 일관성과 가용성을 확보할 수 있도록 공공 데이터베이스에서 신뢰할 수 있는 정보를 중심으로 선정하였다. 각각의 변수는 입지 적합성 점수 산출 과정에서 가중치를 부여받아 최종 점수 계산에 활용되었다. 종속변수는 지역구별 데이터 센터 현재 개수로 설정하였으며, 독립변수는 고립위험, 붕괴시설, 유실위험, 취약방재, 침수위험 등 15개의 변수로 구성된다.

Table 3. Definition of Independent and Dependent Variables for Data Center Suitability Analysis

Category	Variable
Target	Current Number of Data Centers by Region
Feature	Isolation Risk
	Collapse-Prone Facilities
	Loss Risk
	Vulnerable Disaster Prevention
	Flood Risk
	Tsunami Risk
	Earthquake History
	Landslide
	Gas Stations
	LPG Facilities
	Fire Hazard Zones
	Area
	Wind Power Generation
	Dams
	Population

3.2 분석 방법

본 연구는 데이터 센터 최적 입지 선정을 위해 머신러닝 모델과 지리 가중 회귀(Geographically Weighted Regression, GWR) 모델을 활용하였다. 먼저, 머신러닝 모델을 통해 데이터 센터 입지 선정에 영향을 미치는 주요 변수들의 중요도를 분석하였으며, 이후 GWR 모델을 통해 지역별 특성을 반영한 회귀 계수를 산출하였

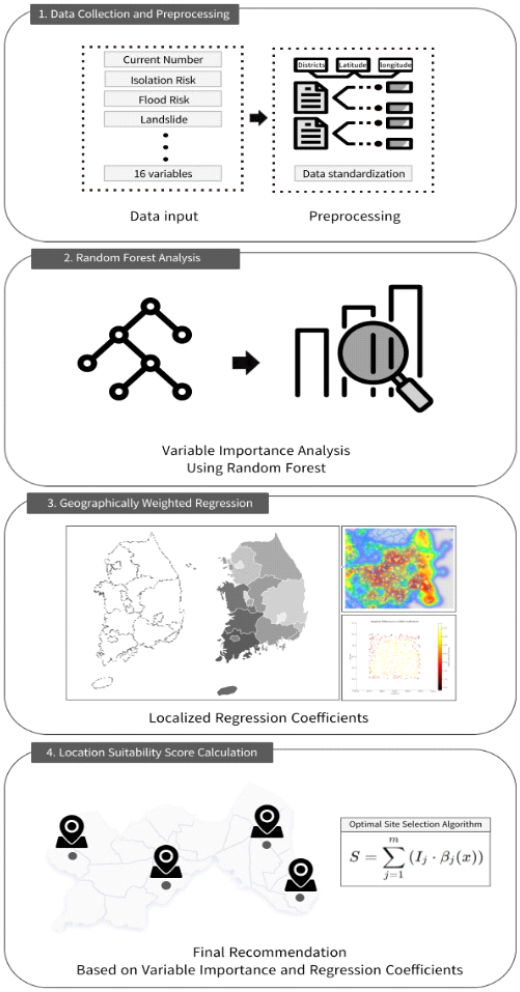

Fig. 1. Overall Process of Data Center Optimal Site Selection

다. 마지막으로, 두 분석 결과를 결합하여 지역별 데이터 센터 입지 적합성을 평가하였다. 본 연구에서 제안하는 방법론의 프로세스는 Figure 1에 나타나 있다. 또한, Table 4에는 본 연구에서 사용된 변수들의 정의가 제시되어 있다.

Table 4. Key Variables for GWR Model Application in Data Center Site Selection

Variable	Definition
y_i	Dependent variables
$\beta_0(u_i, v_i)$	Intercept at the i-th coordinate
$\sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ki}$	Sum of the product of the coefficients of the k-th independent variable and the intercept
ε_i	Error Port
$W_k(x)$	Local weight at location x
d_{x,x_k}	Distance between location x and x_k
b	Bandwidth
$\beta_j(x)$	Regression coefficients by region for variable j
X_j	Value of the variable j

3.2.1 머신러닝 기반 데이터 입지 선정 요인 중요도 분석

본 연구에서는 데이터 입지 선정 요인 중요도 분석을 위해 의사결정나무(Decision Tree), XGBoost, 랜덤 포레스트의 세 가지 머신러닝 기반 회귀 모형들을 활용하였다. 의사결정나무는 데이터를 분리하는 규칙을 통해 예측적 분류 또는 회귀를 수행하는 모델링 기법으로, 각 변수들로 이루어진 트리의 노드들을 순차적으로 살펴보며 분류 또는 회귀를 결정한다[16]. XGBoost는 앙상블 알고리즘 중 하나로, 초기 모델의 예측 오차를 보완하면서 순차적으로 학습을 진행하는 부스팅 방식을 활용한다[17]. 랜덤 포레스트는 다수의 의사결정나무를 앙상블 학습하여 그 결과를 종합하는 모델로, 과적합을 줄이고 예측 성능을 향상시키는데 효과적이다[18].

본 연구에서는 현재 데이터 센터 위치를 모델링하기 위해 전체 데이터셋을 7:3 비율로 훈련 세트와 테스트 세트로 무작위 분할하였다. 모델의 안정성과 일반화 성능을 평가하기 위해 5-fold 교차 검증을 수행하였으며, GridSearchCV를 사용하여 각 모델의 주요 하이퍼파라미터를 최적화하였다. 모델 평가는 Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2)를 기준으로 수행하였으며, 최적의 모델을 선정하여 변수 중요도 분석에 활용하였다. 선정된 모델의 변수 중요도는 GWR 모델의 공간적 특성을 반영한 회귀계수와 결합하여 위치 추천 점수를 산출하는데 활용된다.

3.2.2 GWR 기반 회귀 계수 추정

GWR 모형은 공간적 이질성을 고려하여 회귀 분석을 수행하는 방법으로, 공간적 위치에 따른 변수들의 영향력 변화를 분석할 수 있어 지역별 특성을 고려한 회귀계수를 산출할 수 있다[19]. 전체 데이터에 대해 단일 회귀식을 사용하는 전역적 회귀 모델과 달리, GWR은 각 공간 지점에서의 로컬 회귀 계수를 추정한다. GWR의 기본적인 회귀식은 식(1)과 같다.

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ki} + \varepsilon_i, \quad (1)$$

본 연구에서는 GWR을 활용하여 데이터 센터 입지 선정에 영향을 미치는 변수들의 공간적 상관성과 지역적 특성을 분석하였다. GWR 모형을 적용하기 위해, 먼저 각 시군구에 대한 공간적 가중치 행렬을 계산하였다. 이를 위해 사용된 가중치 행렬은 커널 기반 방식으로, 각 시군구 간의 거리를 고려하여 가까운 지역일수록 더 높은 가중치를 부여하였다. 대역폭값은 최적화 과정에서 결정되었으며, 이를 통해 각 지역 간의 공간적 차이를 적절히 반영하였다. 예를 들어, 침수 위험이나 지진 위험과 같은 변수는 특정 지역에서는 상대적으로 높은 영향을 미치는 반면, 다른 지역에서는 중요도가 낮을 수 있다. 이러한 지역적 차이를 무시할 경우, 입지 선정 모델은 특정 지역의 특성을 제대로 반영하지 못할 가능성이 있다. 따라서 GWR은 각 지역에서의 변수별 회귀계수를 독립적으로 계산함으로써, 지역적 특수성을 반영하고 분석의 유용성을 높인다[19].

GWR 모델에서는 국소적 회귀를 수행하기 위해 커널 기반 공간적 가중치 행렬 W를 사용한다. 이 가중치는 분석 지점에서 가까운 데이터일수록 더 큰 가중치를 부여하고, 멀어질수록 가중치를 줄이는 방식으로 계산된다. GWR의 공간적 가중치 계산 공식은 아래 식(2)와 같다[20].

$$W_k(x) = \frac{\exp(-\frac{d_{x,x_k}^2}{2b^2})}{\sum_{i=1}^{n} \exp(-\frac{d_{x,x_i}^2}{2b^2})} \quad (2)$$

계산된 지역 가중치를 활용한 GWR 회귀 계수는 식 (3) 과 같이 추정된다.

$$\beta_j(x) = \sum_{k=1}^{m} W_k(x) \bullet X_j \quad (3)$$

이 과정에서 각 지역의 고유한 특성이 데이터 센터 입지에 미치는 영향을 정량적으로 평가한다.

3.2.3 데이터 센터 최적 입지 선정 점수

데이터 센터 입지 선정의 최종 점수는 머신러닝 모델의 변수 중요도와 GWR 회귀 계수, 그리고 가중치를 결합하여 계산한다. 최종 점수 계산 공식은 아래의 식 (4)와 같다.

$$S = \sum_{j=1}^{m} (I_j \bullet \beta_j(x)) \quad (4)$$

S : 지역별 최종 점수

I_j : 머신러닝 모델을 통해 산출된 j번째 변수의 중요도

$\beta_j(x)$: GWR을 통해 산출된 변수 j번째 변수의 지역별 회귀 계수

Ⅳ. 분석 결과

4.1 데이터 입지 선정 요인 중요도 분석 결과

본 연구에서는 머신러닝 기반 회귀 모형을 활용하여 지역별 데이터 센터 개수에 영향을 주는 주요 변인들의 중요도를 추출하였다. 최종 머신러닝 모델 선정을 위해 동일한 데이터 실험 환경 하에 의사결정나무, XGBoost, 랜덤 포레스트의 세 가지 머신러닝 모형들의 성능을 비교 분석하였다. 각 모델은 모두 그리드 탐색 기법을 사용하여 최적의 하이퍼파라미터 조합을 탐색하였으며, 5-fold 교차 검증을 수행하여 모델의 안정성을 평가하였다. 의사결정나무의 경우, 최대 트리 깊이는 10, 리프 노드 최소 샘플 수는 5로 설정되었다. XGBoost는 학습률0.1, 최대 트리 깊이 6, 트리 수는 100으로 설정되었다. 랜덤 포레스트 모델은 트리 수 200, 최대 트리 깊이는 15로 설정되었다. 또한, 각 분할에서 최소 샘플 수는 4로, 리프 노드에서 최소 샘플 수는 2로 설정되었다. 이와 같은 하이퍼파라미터 최적화를 통해 모델의 예측 성능을 향상시키고자 하였다. 본 연구는 Intel i9 프로세서와 32GB RAM의 CPU 환경에서 수행되었으며, Python 3.9를 기반으로 Scikit-learn과 XGBoost 라이브러리를 활용하였다. 각 모델의 최종적인 성능은 Table 5와 같다.

Table 5의 분석 결과, 랜덤 포레스트는 가장 낮은 RMSE 값을 보였으며, XGBoost 대비 약 19.2%, 의사결정나무 대비 약 33.3%의 RMSE 감소를 달성하였다. 또한 지표에서도 랜덤 포레스트가 가장 우수한 성능을 보였다. 이를 통해, 데이터 센터의 분포 밀집 정도를 예측하는 최종 모델로 랜덤 포레스트 모형을 선정하였으며, 데이터 센터 입지 선정을 위해 학습된 모형의 변수 중요도를 추출하였다. 랜덤 포레스트의 변수 중요도 분석 결과는 Figure 2에 나타나 있다.

분석 결과, 자연재해 관련 변수인 지진 이력(0.19),

Table 5. Performance Comparison of Machine Learning Models

Model	RMSE	R^2
Decision Tree	0.1845	0.7234
XGBoost	0.1523	0.7862
RandomForest	0.1230	0.8364

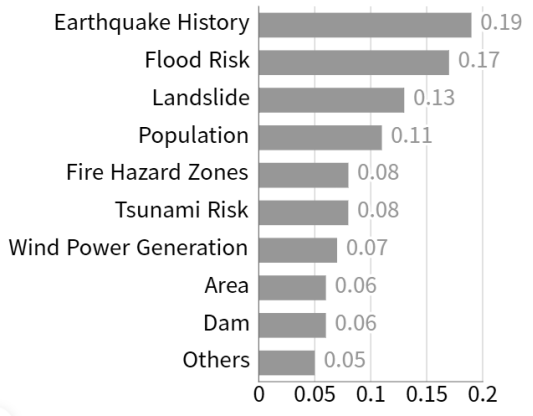

Fig. 2. Feature Importance of Random Forest

침수 위험(0.17), 산사태(0.13)가 가장 높은 중요도를 보였으며, 이는 데이터 센터 입지 선정에 있어 자연재해 요인이 가장 중요한 고려사항임을 시사한다. 특히 이 세 변수의 중요도 합이 0.49로 자연재해 요인이 입지 선정의 핵심 요소임을 보여준다.

다음으로 인구수(0.11)가 높은 중요도를 나타냈는데, 이는 인프라 조건이 입지 선정에 상당한 영향을 미치는 것을 보여준다. 시설 조건과 관련된 변수인 화재 위험 구역(0.08), 풍력발전(0.07), 면적(0.06), 댐(0.06)은 중간 정도의 중요도를 보였다. 이러한 결과는 데이터 센터의 안정적인 운영을 위해서는 자연재해로부터의 안전성이 가장 중요하며, 그 다음으로 인프라 여건이 중요한 요소임을 보여준다.

4.2 GWR 분석 결과

본 연구에서는 GWR을 활용하여 데이터 센터 입지 선정에 영향을 미치는 변수들의 최종 회귀 계수를 추정하였다. GWR 모형의 성능을 평가하기 위해 Akaike Information Criterion (AIC), R^2, RMSE의 지표를 산출하였으며, 그 결과는 Table 6와 같다.

GWR 모델의 성능 평가 결과, AIC 값은 1245.32로 나타나 모델이 데이터를 효과적으로 설명하면서도 복잡성을 최소화했음을 시사한다. Adjusted R^2값은 0.8230으로, GWR 모델이 데이터 변동성의 약 82.3%를 설명할 수 있음을 보여주며, 높은 설명력을 가진 모델임을 입증한다. 또한, RMSE 값이 0.0912로 매우 낮게 나타나 예측값과 실제값 간의 오차가 작아 모델의 예측력이 높음을 확인할 수 있다.

Table 6. Performance Evaluation of GWR Model

Metric	Value
AIC	1245.32
Adjusted R^2	0.8230
RMSE	0.0912

GWR을 통해 계산된 계수값들은 지역별로 유의미한 차이를 보였으며, 일부 지역에 대한 결과 예시는 Table 7에 나타나 있다. 이는 지역적 특성을 반영한 데이터 센터 입지 선정에 있어 중요한 기준점으로 활용될 수 있음을 나타낸다.

대구광역시 북구의 경우, GWR 분석 결과 지진 위험이 데이터 센터 입지 선정에 있어 가장 중요한 양의 영향을 미치는 변수로 나타났다(β=1.35). 이는 대구광역시 북구의 지리적 위치와 지형적 특성이 지진 위험을 상대적으로 더 높게 반영하며, 해당 지역에서의 데이터 센터 운영 안정성에 중요한 고려 사항임을 의미한다. 반면, 침수 위험은 대구광역시 북구에서 데이터 센터 입지 선정에 있어 음의 영향을 미치는 것으로 나타났다(β=−0.15). 이는 침수 위험이 상대적으로 낮거나, 데이터 센터 입지 결정 과정에서 덜 중요한 요소로 간주될 가능성을 시사한다.

경상남도 김해시의 GWR 분석 결과를 보아, 지진

Table 7. Regional GWR Regression Coefficients

Variable	Regression Coefficient $\beta_j(x)$	
	Buk-gu, Daegu	Gimhae, Gyeongsangnam-do
Isolation Risk	-0.01	-0.02
Collapse-Prone Facilities	-0.02	-0.03
Loss Risk	-0.01	-0.01
Vulnerable Disaster Prevention	-0.01	-0.10
Flood Risk	-0.15	-0.10
Tsunami Risk	0.17	0.05
Earthquake History	1.35	1.02
Landslide	1.18	0.15
Gas Stations	0.02	0.03
LPG Facilities	-0.01	-0.05
Fire Hazard Zones	0.51	0.08
Area	0.08	0.12
Wind Power Generation	0.20	0.18
Dams	0.05	0.06
Population	0.30	0.20

위험(β=1.02)과 인구 밀도(β=0.20)가 주요한 양의 영향을 미치는 변수로 나타났다. 이는 김해시의 지리적 특성과 도시적 특성이 데이터 센터 운영에 중요한 요소로 작용하며, 각각의 변수가 데이터 센터 입지 결정 과정에서 고려해야 할 핵심 요인임을 의미한다.

4.3 데이터 센터 입지 적합성 점수 산정 결과

본 연구에서는 지역별 데이터 센터의 최종 입지 적합성 점수를 머신러닝 모델 기반 중요도와 GWR 계수를 결합하여 산출하였다. 최종 점수는 각 지역의 데이터 특성을 기반으로 계산되었으며, Table 8에 변수별 랜덤 포레스트 중요도와 GWR 회귀계수를 곱하여 산출된 입지 적합성 점수 계산 예시가 제시되어 있다.

Table 9은 이러한 계산 방식을 통해 도출된 전국 시군구별 데이터 센터 입지 적합성 점수를 기준으로, 현재 데이터 센터가 위치하지 않은 지역 중 상위 15개 지역을 선정하여 제시하였다. 또한, Figure 3은 이를 기반으로 상위 지역을 지도에 시각화한 것 이다. 분석 결과, 충청남도 천안시가 0.682점으로 가장 높은 입지 적합성 점수를 기록했으며, 경상남도 김해시(0.645)와 대구광역시 북구(0.494)가 높은 점수를 보였다.

상위권에 위치한 지역들은 대체로 자연재해 위험이 상대적으로 낮고, 인프라 조건이 우수하며, 전력 공급이 안정적인 특징을 보였다. 특히 천안시와 김해시는 Figure 2에서 제시된 주요 변수들 중 지진이력, 침수위험, 산사태 등 자연재해 관련 변수에서 안정적인 값을 보이면서도, 인구수와 같은 인프라 조건에서도 높은 점수를 획득했다. 특히, 천안시는 경제적 중심지로서 인프라와 확장성 면에서 데이터 센터 입지로 최적의 조건을 제공한다. 김해시는 데이터 센터 운영에서 수요와 확장

Table 8. Example of Data Center Suitability Score Calculation

Variable	Buk-gu, Daegu		
	Random Forest Importance I_j	GWR Regression Coefficient $\beta_j(x)$	Final Score S
Isolation Risk	0.001	-0.01	-0.00001
Collapse-Prone Facilities	0.0033	-0.02	-0.000066
Loss Risk	0.0137	-0.01	-0.000137
Vulnerable Disaster Prevention	0.012	-0.01	-0.00012
Flood Risk	0.17	-0.15	-0.0255
Tsunami Risk	0.08	0.17	0.0136
Earthquake History	0.19	1.35	0.2565
Landslide	0.13	1.18	0.1534
Gas Stations	0.01	0.02	0.0002
LPG Facilities	0.01	-0.01	-0.0001
Fire Hazard Zones	0.08	0.51	0.0408
Area	0.06	0.08	0.0048
Wind Power Generation	0.07	0.20	0.0140
Dams	0.06	0.05	0.0030
Population	0.11	0.30	0.0330
Sum	1	-	0.4935

Table 9. Top 15 Recommended Locations for New Data Center Construction

Rank	Region	Location Score
1	Cheonan-si, Chungcheongnam-do	0.682
2	Gimhae-si, Gyeongsangnam-do	0.645
3	Buk-gu, Daegu Metropolitan City	0.494
4	Uiwang-si, Gyeonggi-do	0.482
5	Ulju-gun, Ulsan Metropolitan City	0.475
6	Dongnae-gu, Busan Metropolitan City	0.463
7	Yeongcheon-si, Gyeongsangbuk-do	0.458
8	Suncheon-si, Jeollanam-do	0.442
9	Yesan-gun, Chungcheongnam-do	0.435
10	Yeongdo-gu, Busan Metropolitan City	0.428
11	Seongju-gun, Gyeongsangbuk-do	0.412
12	Sokcho-si, Gangwon Special Self-Governing Province	0.395
13	Goseong-gun, Gyeongsangnam-do	0.385
14	Uljin-gun, Gyeongsangbuk-do	0.374
15	Yeongyang-gun, Jeollanam-do	0.362

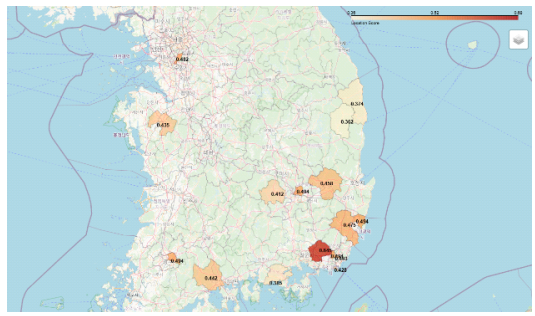

Fig. 3. GIS Visualization for Top-15 Recommended Locations for Data Centers Map

성이 높은 지역으로 평가되었다. 김해시는 높은 사회적 인프라와 안정적인 고객 수요 기반을 제공할 수 있는 지역으로, 데이터 센터 입지로서 긍정적인 가능성을 보인다. 울산광역시 울주군(0.475)은 산업 인프라와 넓은 가용 면적에서 강점을 보였다. 강원도 고성군의 최종 점수는 0.385의 점수를 기록하였다. 이는 지진 위험과 인구 밀도의 상대적으로 낮은 영향으로 인해 점수가 비교적 제한된 것으로 보인다. 그러나 이러한 지역들도 적절한 인프라 구축과 지원 정책을 통해 데이터 센터 입지로서의 잠재력을 발현할 수 있을 것으로 기대된다.

추가적으로, 본 연구에서는 선정된 데이터 센터 입지의 우수성을 입지 위험성의 측면에서 검증하기 위해 선행연구에서 많이 활용되는 다기준 세트 커버링 모델과의 주변 환경 위험 점수 비교를 진행하였다[15]. 다기준 세트 커버링 모델은 정해진 기준을 만족하는 최소 개수의 시설을 선정하는 최적화 문제를 선형 계획법 기반으로 풀어내는 기법이다. 본 연구에서는 다른 변수들을 제약조건으로 가지면서 기준을 풍력 및 인프라 최대화와 위험 최소화로 산정하여 분석 후 제안하는 방법과의 비교를 진행하였다. Figure 4는 두 모델의 최종 선정 입지들 간 위험도 차이 비교 결과를 시각화 한 것이다.

Figure 4를 살펴보면, 다기준 세트 커버링 모델로부터 선정된 입지들의 경우 본 연구에서 선정된 상위 추천 입지 대비 위험도가 높은 것을 확인할 수 있다. 이는 본 연구에서 제안하는 공간적 정보 및 머신러닝 기반의 변수 특성 정보를 복합적으로 반영하는 것이 선정 입지의 위험 최소화 측면에서도 효과적임을 나타낸다.

최종 점수 계산 결과, 지역별로 데이터 센터 입지 선정에 영향을 미치는 요인들이 상이하게 작용함을 확인할 수 있었다. 이러한 분석 결과는 데이터 센터의 지

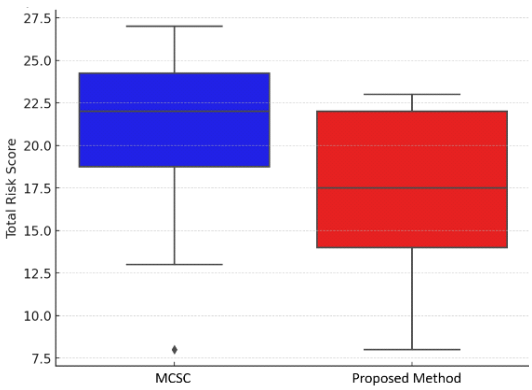

Fig. 4. Risk Comparison Results of Recommended Locations for Data Centers. MCSC is Multi-Criteria Set Covering

방 분산이 에너지 수급 안정성과 지역 균형 발전을 고려한 현실적인 대안이 될 수 있음을 시사하며, 수도권 집중 완화와 지역 간 인프라 격차 해소를 동시에 달성할 가능성을 제시한다. 또한, 각 지역의 지형적·환경적 특성을 반영한 맞춤형 데이터 센터 개발 전략의 필요성을 강조하며, 향후 데이터 센터 입지 선정 시 본 연구에서 도출된 순위 및 주요 입지 요인을 고려한 전략적 의사결정이 가능할 것으로 기대된다.

특히, 천안시와 김해시는 자연재해 위험이 상대적으로 낮고 인구 기반이 탄탄하며, 전력 공급 및 사회적 인프라 조건에서도 우수한 평가를 받아 데이터 센터 구축에 적합한 지역으로 도출되었다. 대구광역시 북구는 자연재해 요인에서 다소 높은 위험을 보였으나, 이를 감안한 안정적인 운영 환경이 구축될 경우 데이터 센터 입지로 충분히 고려할 만한 지역임을 시사한다. 또한, 울산광역시 울주군과 강원도 고성군은 넓은 가용 면적과 인프라 조건에서 강점을 가지지만, 자연재해 위험 요소가 일부 존재하는 만큼 추가적인 안정성 검토가 필요한 지역으로 나타났다. 결과적으로, 본 연구에서 도출된 데이터 센터 적합 지역 분석은 지역적 특성과 주요 입지 요인 간의 관계를 정량적으로 평가하는 데 기여하며, 추후 데이터 센터 입지 선정 및 정책 수립 시 참고할 수 있는 기초 자료를 제공할 수 있을 것으로 기대된다.

V. 결 론

본 연구는 데이터 센터 입지 선정의 효율성과 지속가능성을 높이기 위해 머신러닝(Random Forest)과 지리 가중 회귀(GWR)를 활용한 입지 적합성 평가 방법론을 제안하였다. 랜덤 포레스트를 통해 데이터 센터 입지에 영향을 미치는 주요 변수(자연재해, 인프라, 전력 공급 안정성 등)의 중요도를 분석하고, GWR을 적용하여 지역별 특성을 반영한 회귀 계수를 산출하였다. 이를 바탕으로 데이터 센터의 최적 입지를 평가한 결과, 천안, 김해, 대구 등 일부 지방 지역이 수도권 대비 높은 입지 적합성을 보이며, 자연재해 위험이 낮고 인프라 접근성이 우수한 지역이 데이터 센터 운영에 유리함을 확인하였다.

본 연구의 학술적 기여는 다음과 같다. 첫째, 머신러닝과 공간 회귀 분석을 결합한 정량적 입지 평가 방법론을 제시하여 기존 연구의 한계를 보완하고, 복잡하지 않은 모형을 활용하여 분석 결과를 효율적이고 효과적으로 전달할 수 있는 접근법을 제시하였다. 둘째, 자연재해, 인프라, 전력 공급 등 주요 요인 간의 상호작용을 반영하여, 현재 가용한 데이터 환경에서 도출할 수 있는

최적의 입지 평가 결과를 다양한 분석과 함께 제공하였다. 셋째, 본 연구는 새로운 데이터나 변수를 추가하여 확장할 수 있는 유연한 프레임워크를 제안하였으며, 데이터 센터뿐만 아니라 스마트시티, 클라우드 인프라 등 다양한 기반 시설의 입지 선정 연구에도 적용할 수 있는 초석을 마련하였다.

본 연구는 현실적인 분석 환경 내에서 데이터 센터 입지 선정에 대한 분석을 수행하였으나, 몇 가지 한계점이 존재하여 이에 대한 보완 계획 및 향후 연구를 제안하고자 한다. 첫째, 본 연구에서 활용한 데이터는 특정 기간 범위의 국내 데이터로 한정되어 있어 추후 글로벌 데이터로 확장하여 적용할 수 있다. 둘째, 본 연구에서는 데이터 센터 운영에 영향을 미칠 수 있는 에너지 비용, 탄소 배출, 지역 사회의 수용성 등의 경제적·환경적 요인의 경우 지역단위 데이터 수집의 어려움으로 인해 모형에 반영되지 않았다. 향후 연구에서는 앞으로 다양한 관련 데이터베이스가 점진적으로 구축되고 활용될 가능성이 높아짐에 따라, 추가적인 요인들을 포함하는 보다 정교한 데이터 센터 입지 분석을 위한 데이터 구축-정제-분석을 전범위적으로 고려하는 연구의 수행이 필요할 것으로 사료된다. 또한, 기존 연구들과의 차별성을 정량적·정성적으로 검증할 수 있는 추가적인 분석 방법을 마련할 필요가 있으며, 데이터 센터 입지 선정의 현실성을 높이기 위해 다양한 공간적·사회적 요소를 포함하는 확장된 변수와 데이터 활용 방안을 지속적으로 모색하고자 한다.

References

[1] J. G. Koomey, "*Growth in data center electricity use 2005 to 2010*," *Analytics Press*, 2011, Retrieved from https://www.alejandrobarros.com/wp-content/uploads/old/4363/Growth_in_Data_Center_Electricity_use_2005_to_2010.pdf

[2] N. Jones, "How to stop data centres from gobbling up the world's electricity," *Nature*, vol. 561, no. 7722, pp. 163-166, 2018. (https://doi.org/10.1038/d41586-018-06610-y)

[3] International Energy Agency (IEA), *Electricity 2024: Analysis and forecast to 2026*, International Energy Agency, 2024, Retrieved from https://www.iea.org/reports/electricity-2024.

[4] S.-H. Lee, S.-J. Moon, J.-H. Kim, S.-Y. Shin, and Y.-J. Choi, "Strategies for implementing green data centers in the public sector," *J. Korean Soc. Inf. Process.*, vol. 16, no. 6, pp. 54-65, 2009.

[5] K.-W. Kim and C.-S. Kim, "A study on cloud data center location selection considering disaster information," *J. KIICE*, vol. 16, no. 12, pp. 2575-2580, 2012.

[6] Korea Economic Daily, "*SK C&C data center fire halts KakaoTalk service*," 2022, Retrieved from http://www.hankyung.com

[7] Microsoft, "*Project natick: Microsoft's moonshot to make the world's data centers more sustainable*," 2020, Retrieved from https://natick.research.microsoft.com

[8] J. Song, "Improving data center sustainability and achieving ESG goals," *Korea Data Center Energy Efficiency Association*, 2022.

[9] J. Lee, M. Hong, and S. Lee, "Application of overseas design standards for global data centers," in *Proc. Korean Soc. Mechanical Eng. Winter Conf.*, pp. 174-178, 2021.

[10] J.-G. Cho and K.-J. Lee, "A formal study of skyscraper data center in city for lowering cooling load," in *Proc. Architectural Inst. Korea Autumn Conf.*, vol. 41, no. 2, pp. 55-58, 2021.

[11] M.-T. Kim, "Finding the way between domestic data center industry, AI, and ESG," *Hana Financial Management Res. Inst.*, Mar. 2024.

[12] J.-H. Kan, S.-H. Nam, Y.-J. Seo, E.-J. Jeon, and K.-C. Lee, "Proposal for site selection of green corridors and green spaces for alleviating the urban heat island effect in Seoul: Focused on heat island scale analysis using clustering and random forest importance," *J. Digital Convergence*, vol. 6, no. 1, pp. 27-38, 2019.

[13] H. Kim, S.-H. Cho, and S.-K. Lee, "Strategic design of green data centers: A case study of public sector data centers in Korea," *J. Korea Inf. Technol.*, vol. 10, no. 4, pp. 143-152, 2012.

[14] A. Covas, F. Pires, and J. Almeida, "Multi-criteria decision analysis for sustainable data center location," *J. Econ. Anal. & Policy*, vol.

3, no. 1, Article 21, 2012. (http://dx.doi.org/10.1111/j.1475-3995.2012.00874.x)

[15] M. Kheybari, A. Arbabian, and S. Zolfaghari, “Multi-criteria set covering decision for sustainable data center location,” *Econ. Lett.*, vol. 86, no. 1, pp. 107-112, Jan. 2020. (https://doi.org/10.1142/S0219622020500157)

[16] D.-H. Lee and J.-H. Yang, “Exploring trends in domestic research on decision tree analysis,” *J. KDAS*, vol. 26, no. 5, pp. 1375-1393, 2024.

[17] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proc 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785-794, Aug. 2016. (https://doi.org/10.1145/2939672.2939785)

[18] R. Iranzad and X. Liu, “A review of random forest-based feature selection methods for data science education and applications,” *Int. J. Data Sci. and Analytics*, pp. 1-15, 2024. (https://doi.org/10.1007/s41060-024-00314-7)

[19] A. S. Fotheringham, et al., “Exploring spatial context: A comprehensive bibliography of GWR and MGWR,” *arXiv preprint arXiv: 2404.16209*, 2024.

[20] O. Raza, M. A. Mansournia, A. R. Foroushani, and K. Holakouie-Naieni, “Geographically weighted regression analysis: A statistical method to account for Spatial Heterogeneity,” *Archives of Iranian Medicine*, vol. 22, no. 3, pp. 155-160, 2019. (https://doi.org/10.1016/j.archir.2019.03.015)

이 우 진 (Woojin Lee)

2017년 2월 : 영남대학교 국제통상학과 상경학사

2023년~현재 : 경북대학교 데이터사이언스대학원 석사

<관심분야> 인공지능 응용, 그래프 기반 딥러닝, 추천 시스템, 대규모 언어 모델

[ORCID:0009-0002-5664-983X]

김 민 영 (Minyoung Kim)

2023년 2월 : 영남대학교 정보통신공학과 졸업

2025년 2월 : 경북대학교 데이터사이언스학과 석사 졸업

<관심분야> 자연어 처리, 추천 시스템, 인공지능 응용

[ORCID:0009-0006-6032-7621]

윤 선 영 (Sunyoung Yoon)

2022년 2월 : 인제대학교 통계학과 졸업

2025년 2월 : 경북대학교 데이터사이언스학과 석사 졸업

<관심분야> 인공지능 응용, 데이터 마이닝

[ORCID:0009-0008-9021-4788]

김 수 현 (Suhyeon Kim)

2018년 2월 : 부산대학교 대기환경과학과/통계학과 이학사

2020년 2월 : 울산과학기술원 융합경영대학원 비즈니스분석 이학석사

2023년 2월 : 울산과학기술원 산업공학과 공학박사

2023년~현재 : 경북대학교 데이터사이언스대학원 교수

<관심분야> 데이터 마이닝, 인공지능 응용, 모빌리티 데이터 분석, 연합학습

[ORCID:0000-0001-6475-2461]