JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 8, pp. 165-170, August 2024
https://doi.org/10.9708/jksci.2024.29.08.165

Verification of the Suitability of Fine Dust and Air Quality Management Systems Based on Artificial Intelligence Evaluation Models

Heungsup Sim*

*Professor, Computer & Military Department, Dongyang University, Yeongju City, Korea

[Abstract]

This study aims to verify the accuracy of the air quality management system in Yangju City using an artificial intelligence (AI) evaluation model. The consistency and reliability of fine dust data were assessed by comparing public data from the Ministry of Environment with data from Yangju City's air quality management system. To this end, we analyzed the completeness, uniqueness, validity, consistency, accuracy, and integrity of the data. Exploratory statistical analysis was employed to compare data consistency. The results of the AI-based data quality index evaluation revealed no statistically significant differences between the two datasets. Among AI-based algorithms, the random forest model demonstrated the highest predictive accuracy, with its performance evaluated through ROC curves and AUC. Notably, the random forest model was identified as a valuable tool for optimizing the air quality management system. This study confirms that the reliability and suitability of fine dust data can be effectively assessed using AI-based model performance evaluation, contributing to the advancement of air quality management strategies.

▸ **Key words**: Data quality, data preprocessing, random forwarding, machine learning, model performance evaluation

[요 약]

본 연구는 인공지능 평가 모델을 활용하여 양주시의 대기질 관리 시스템의 정확성을 검증하는 데 목적이 있다. 환경부 미세먼지 공공 데이터와 양주시 대기질 관리 시스템 데이터를 비교하여 미세먼지 데이터의 정합성과 신뢰성을 평가하였다, 이를 위해 데이터의 완전성, 유일성, 유효성, 일관성, 정확성, 무결성을 분석하였다. 데이터의 정합성을 비교하기 위해 탐색적 통계 분석을 활용하였다. 분석 결과, AI 기반 데이터 품질 지수 평가 결과, 두 데이터 세트 간에 통계적으로 유의미한 차이가 없음을 확인하였다. AI 기반 알고리즘 중 랜덤 포레스트 모델이 가장 높은 예측 정확도를 보였으며, ROC 커브와 AUC를 통해 예측 성능을 평가하였다. 특히, 랜덤 포레스트 모델은 대기질 관리 시스템의 최적화에 유용한 모델로 확인되었으며, 미세먼지 데이터의 신뢰성과 적합성을 AI 기반 모델 성능 평가로 활용할 수 있음을 확인하였다.

▸ **주제어**: 데이터품질, 데이터전처리, 랜덤포래스트, 머신러닝, 모델성능평가

• First Author: Heungsup Sim, Corresponding Author: Heungsup Sim
*Heungsup Sim (mylee911@naver.com), Computer & Military Department, Dongyang University
• Received: 2024. 07. 02, Revised: 2024. 08. 13, Accepted: 2024. 08. 16.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 연구에서는 더욱 고도화된 AI 모델, 특히 딥러닝 모델을 사용하여 미세먼지 예측의 정확도를 높이는 방향으로 발전하고 있다. 기존의 랜덤 포레스트, 서포트 벡터 머신과 함께 CNN, RNN등 딥러닝 모델의 활용이 증가하고 있다[1]. 검증된 미세먼지 데이터는 실시간으로 대기질을 예측하고 경고하는 시스템은 빠른 데이터 처리와 정확한 예측을 위해 AI 모델과 클라우드 컴퓨팅, 엣지 컴퓨팅 기술을 통합하고 있다[2].

미세먼지 대기오염을 이해하고 오염률을 줄이기 위한 모델과 방법 설계에 관한 IoT와 AI 기반 하이브리드 모델로 대기질 지수(AQI) 예측 기술에 대한 정확도 및 검증 요구가 커지고 있다[3].

본 연구의 주요 방향은 다음과 같다, 미세 먼지 종합포탈의 공공 데이터와 양주시 미세먼지 환경 모니터링 서비스의 위치 기반 RAW 데이터를 비교하여 대기질 관리 시스템의 신뢰성과 일관성을 검증한다. 미세먼지 측정 장치 위치의 적합성을 확보하고 분석 결과를 도출한다. 이는 측정 데이터의 대표성과 신뢰성을 높이는 데 기여할 것이다. 다양한 AI 기반 알고리즘을 적용하여 모델 성능 평가를 수행하고, 예측 성능이 가장 높은 최적의 알고리즘을 도출한다. 이를 통해 미세먼지 예측의 정확도를 높이고 효과적인 대기질 관리 방안을 제시하고자 한다. 데이터 품질 지수 평가, 독립표본 T-검정, AI 기반 모델 성능 평가 등 다양한 방법론을 적용하여 연구의 타당성과 신뢰성을 확보한다.

본 연구의 결과는 양주시뿐만 아니라 다른 지역의 대기질 관리 시스템 개선에도 적용 가능한 평가 모델을 제시하고자 한다. 품질 지수 평가를 통한 데이터 신뢰성은 품질 지표의 완전성, 고유성, 타당성, 일관성, 정확성, 무결성이 포함되며 이를 정량화하고 평가하여 데이터의 신뢰성을 결정한다[1].

독립표본 T-검정을 이용한 두 그룹 비교 본 연구에서는 AI 기반으로 데이터 도구를 이용하여 R과 Python의 라이브러리, 패키지 등 다양한 도구로 분석한 미세먼지 데이터 결과를 비교했다. 도출된 분석 결과를 '오렌지 데이터 마이닝' 과 AI기반 검증 모델의 'KAMP 제조 AI 데이터분석 도구로 AI 플랫폼과 비교하여 결과의 차이를 검증했다.

AI 기반 모델 성능 평가를 통한 최적 알고리즘 적합성은 실증 데이터를 활용하여 AI 기반 미세먼지 예측_분류 알고리즘의 적합성 연구를 진행 했다. 각 알고리즘의 특성을 토대로 최적의 알고리즘의 적합성을 평가하고, 다수결의 법칙에 따라 가장 성능이 좋은 모델을 추천하였다[2].

II. Preliminaries

1. Related works

1.1 Data rQuality Index Evaluation

미세먼지 데이터를 바탕으로 미세 먼지 종합포탈의 공공데이터와 양주시 미세먼지 모니터링 서비스의 데이터의 품질지수 평가 분석을 데이터 품질 지수별로 각각 분석하여 비교 평가 하였다.

두 지점간의 데이터의 일치성과 품질지수를 통한 정합성 검증을 하였다.

표1은 데이터 품질 지수로는 총 6가지를 평가항목의 계산공식을 나타낸다.

완전성(Completeness) 필수항목에 누락이 없어야 한다. 유일성(Uniqueness)데이터 항목은 유일해야 하며 중복되어서는 안 된다.

유효성(Validity) 데이터 항목은 정해진 데이터 유효범위 및 도메인을 충족해야 한다.

일관성(Consistency) 데이터가 지켜야 할 구조, 값, 표현되는 형태가 일관되게 정의 되고, 서로 일치해야 한다. 정확성(Accuracy) 실제 존재하는 객체의 표현 값이 정확하게 반영이 되어야 한다.

무결성(Integrity) 데이터베이스 자료의 오류 없이 변화에 영향을 받지 않고 데이터의 정확성과 일관성, 유효성이 보호되어야 한다[3].

Table 1. Data Quality Application Formula

Sortation	application formula
Completeness	(1-(missing/N)) * 100
Uniqueness	((Only number of data)/N)) * 100
Validity	(Validity Satisfaction Data/N))
Consistency	(consistency satisfaction data/N)
Accuracy	(1-(Accuracy violation data count/N))
Integrity	(1-(The number of non-100% of the uniqueness, validity, and consistency indices/3)) * 100

데이터 품질 지수 평가 연구 결과[그림1]는 공공데이터의 데이터 품질 지수 평가 결과[표2,]데이터 세트는 완전하고 유효하며 일관성과 정확성이 증명된다. 고유성 지수가 낮아 고려되는 열에 고유한 값이 없음을 확인했다.

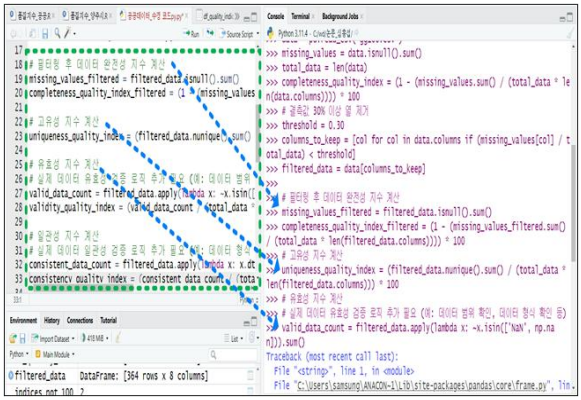

Fig. 1. Calculation of data quality index

Table 2. Data quality figures

Public Data Index	PYTHON	R
Completeness	99.67	100
Uniqueness	32.31	32.31
Validity	100	100
Consistency	100	100
Accuracy	100	100
Integrity	66.67%	66.67

무결성 지수는 낮은 고유성 지수의 영향을 받았으며, 표2는 다른 분석도구로 분석 결과 python과 R의 결과 화면이 거의 동일한 것을 알수 있다.

양주시 모니터링 데이터를 동일한 연구 방법으로 측정한 결과를 보면 표3과 같다. 완전성 품질 지수는 전처리 전과후를 비교하여 초기 데이터 세트의 완전성을 측정하고, 완전성 품질 지수(필터링됨) 누락된 값이 30% 이상인 열을 필터링한 후 데이터 세트의 완전성을 측정하였다.

Table 3. Comparison table of data quality figures

Yangju City Data Index	PYTHON	R
Completeness	99.67	100
Uniqueness	28.87	28.87
Validity	100	100
Consistency	100	100
ccuracy	100	100
Integrity	66.67%	66.67

그림2는 고유성 품질 지수 데이터의 양주시 미세먼지 환경 모니터링 서비스 고유성을 측정하였다.

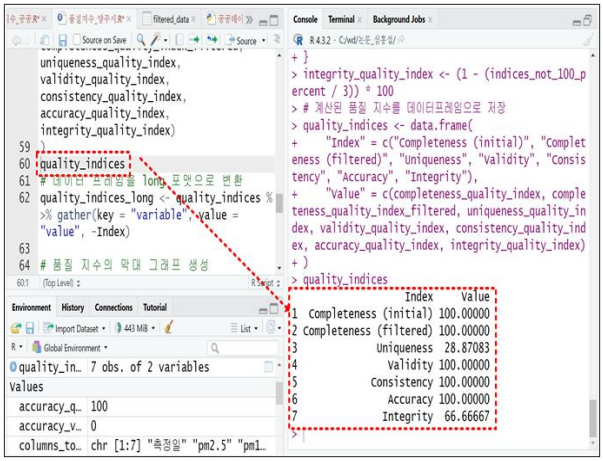

Fig. 2. Final figures of the fine dust data quality index for monitoring in Yangju City.

유효성 품질 지수 모든 데이터가 유효 범위 내에 데이터의 유효성을 측정 후, 일관성 품질 지수 모든 데이터가 일관성이 있다는 가정하에 데이터의 일관성을 측정한다[4]. 정확도 품질 지수 정확도 위반이 없다는 가정 하에 데이터의 정확도를 측정, 무결성 품질 지수 고유성, 타당성, 일관성을 고려하데이터의 전반적인 무결성을 측정하였다.

그림3은, 품질 지수의 막대 그래프 생성을 통하여 연구결과를 비교하여, 미세 먼지 종합포탈의 지수와 양주시 미세먼지 모니터링 서비스 지수를 시각화 표현하여 미세먼지 데이터 품질지수 출력의 양주시 모니링 데이터 결과로 R도구 활용하였다.

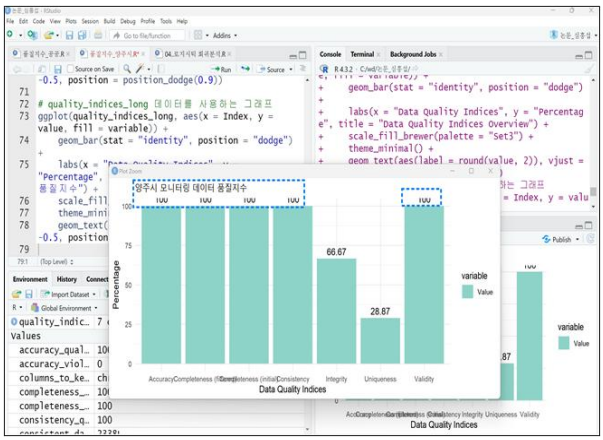

Fig. 3. Fine dust data quality index output_monitoring data_R for both weeks

1.2 Comparison of two groups using independent sample T-test_consistency

독립 표본 T 테스트는 두 그룹의 표본 평균 간의 차이가 통계적으로 유의미한지를 검증하는 통계 검정 방법으로 “미세 먼지 종합포탈의 지수와 양주시 미세먼지 모니터

링 서비스 지수"를 데이터 마이닝 처리 후 파생변수를 만들어서 파생변수의 범주형 데이터 변환 후 미세먼지의 수치를 순서형 범주형 으로 변환한다. 변환된 지표는 등급비교를 통하여 정합성을 비교 했다.

연구 가설 설정은 다음과 같다.

귀무가설 (H0)은 두 그룹의 평균은 차이가 없다.

연구가설 (H1)은 두 그룹의 평균은 서로 다르다.

등분산분석 결과는 F = 0.90537, num df = 363, denom df = 363, p-value = 0.3441로 등분산이다.

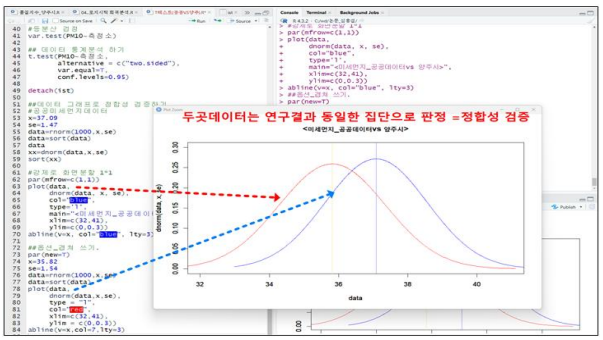

Fig. 4.Consistency Verification Graph_T-TEST

연구 결과 그림4처럼, t = 0.59378, df = 726, p-value = 0.5528로 통계적으로 귀무가설을 채택한다.

1.3 An optimal algorithm suitability study through AI-based model performance evaluation

다양한 AI 모델 비교에 따른 최적의 모델을 추천하고, 추천된 모델을 기반으로 양주시 미세먼지 모니터링 데이터를 실행한다. AI기반의 모델 평가 방법과 평가 모델의 강화 학습을 통한 분류 예측의 연구 결과를 도출하였다. 데이터간의 검증력과 정합성을 바탕으로 데이터 모델의 적합성을 검증하였다[.5]

그림5는 오랜지 데이터마이닝의 지도 학습에는 퓨쳐 값(독립 변수, 특징 또는 예측 변수라고도 함)과 라벨(반응 변수, 종속 변수, 레이블, 타깃 또는 출력 값)을 모두 사용됩니다. 이 기법은 퓨쳐 데이터(설명 변수)와 타깃 데이터(목표 변수) 간의 관계를 이해하고 모델링하는 데 중점을 둔다. 지도 학습의 주요 목표는 이러한 학습된 관계를 기반으로 미래 관찰에 대해 정확한 예측을 하는 것, 특히 인식, 분류, 진단, 예측과 관련된 문제를 해결하는 데 효과적이다[6]. 연구에 활용한 지도 학습 분류 알고리즘에는 K-최근접 이웃, 로지스틱 회귀분석, 인공신경망 분석, 의사결정 트리, 서포트 벡터 머신, 나이브 베이즈, 앙상블 기법(랜덤 포레스트)을 활용한 결과를 도출하였다.[7]

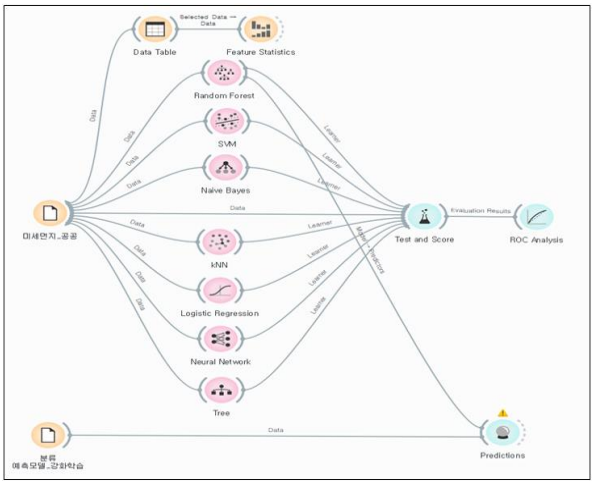

Fig. 5. ML-based optimal algorithmic random forest _ long-range data mining

연구 분석 결과 그림5는 오랜지 데이터 마이닝을 활용한 ML기법의 모델성능평가 알고리즘의 결과 프로세스를 보여주는 그림으로 테스트엔스코어의 버튼 실행결과 그림6의 화면으로 전환된다.

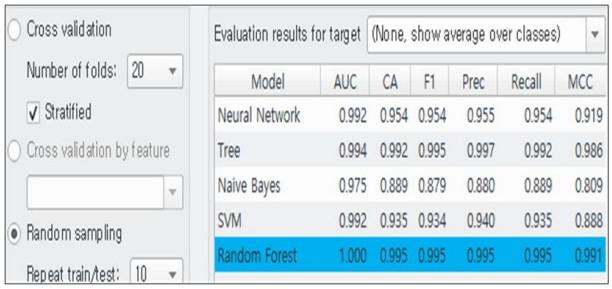

Fig. 6. Model Performance Evaluation Results

그림6은 최종 추천모델의 결과 랜덤포래스트를 활용 하였을때의 모델 성능 평가 지수가 가장 유효한 결과가 도출되어 미세먼지포털의 공공데이터의 최적합 모델로는 "Random Forest " 알고리즘을 결정하였다.

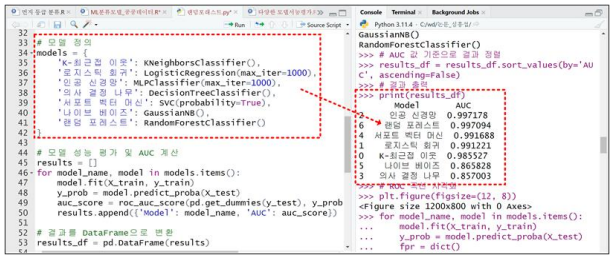

Fig. 7. Model Performance Evaluation Results_R

다른 모델도 성능이 우수하게 나아왔으나, 그림7은 R을 통한 모델 성능 평가 분석에서도 최종 Random Forest의 모델이 우수한 것을 확인 할수 있었다.

III. The Proposed Scheme

Table 4. Process of application of each analysis tool and verification tool

Sortation	application	Analysis area		
R	Analysis	Data quality	T-TEST	ML_ Model Performance Evaluation Random Forest
PYTHON	Analysis	Data quality	T-TEST	
KAMP	Verification			Random Forest
GPT	Verification	Data quality	T-TEST	Model Performance Evaluation Random Forest
CUE:	Code Analysis	Some validation tools for code in the analysis area.		
Orange data mining	Verification			Model Performance Evaluation Random Forest

표4는 분석 목적별 활용한 도구에 따른 분석 영역의 검증도구별 사례를 나열한 표이다. AI LLM모델의 GPT와 카이스트에서 개발한 KAMP의 결과도 Random Forest로 나타냄을 확인하였다. 데이터정합성 연구시, 도구별 수치검증부분으로 활용하여 데이터의 신뢰성과 유효성을 확인하여 활용이 가능함을 확인하였다.

미세먼지 종합포탈의 공공데이터를 RAW 데이터로 활용하여 데이터 품질 지수를 평가하고, 데이터의 신뢰성을 연구에서는 데이터 품질 지수 항목으로는 완전성, 유일성, 유효성, 일관성, 정확성, 무결성을 수치화하여 정량적으로 연구했다. 이러한 데이터 품질 지수 평가 결과, 미세먼지 데이터의 전반적인 신뢰성이 확인되었다[9].

특히, 데이터 전처리 과정을 통해 결측값을 처리하고, 데이터의 일관성을 보장함으로써 신뢰성 있는 데이터셋을 확보할 수 있었다[10].

독립표본 T-테스트를 활용한 두 집단 비교_정합성을 통한 연구 방법론으로는, 독립표본 T-테스트를 활용하여 두 집단 간의 미세먼지 데이터의 정합성을 비교한 결과, 두 집단 간의 평균값에 유의미한 차이가 없다는 결론을 검증했다[11].

AI 기반 최적의 알고리즘 적합성 연구로AI 기반의 미세먼지 예측분류 모델의 성능 평가를 통해 Random Forest 알고리즘을 최적의 알고리즘을 제시했다. 다양한 머신러닝 알고리즘(K-Nearest Neighbors, Logistic Regression, Artificial Neural Network, Decision Tree, Support Vector Machine, Naive Bayes, Random Forest)을 비교 모델 성능평가한 결과, Random Forest 모델이 가장 높은 예측 성능을 보였다. AUC 및 ROC 지표를 통해 모델의 성능을 평가한 결과, Random Forest 모델의 AUC 값이 가장 높게 나타났다.

AI 기반의 미세먼지, 대기질 관리시스템의 적합성을 검증하고, 데이터 품질 지수 평가를 통해 데이터의 신뢰성을 확인했다.

IV. Conclusions

본 연구는 인공지능 평가 모델 기반의 미세먼지 및 대기질 관리시스템의 적합성 검증에 관한 연구로 데이터 품질 평가 지수 분석, 독립표본 T-테스트를 활용한 두 집단 비교_정합성, AI 기반 최적의 알고리즘 적합성 연구라는 연구 절차 방법을 수행 결과 적합성에 타당한 결론을 유추할수 있었다.

Table 5. Research Ideas Based on the Classification Model Evaluation Results

Evaluation	Classification Model Evaluation	Result
Major Goals	Accurately predict the classification rate of pollutants (apply to 7 analysis methods).	Random Forest: Excellent model performance for PM10 level classification.
Common Measurement Items	Accuracy / Precision / Recall / F1 Score ROC curve area (AUC-ROC)	AUC=1.00 F1=0.98 ROC area = 0.98 CA=0.98 PRE=0.98 RECALL=0.98
Predictive Interpretation	Predictions for each member are necessary because each class is predicted independently (e.g., logistic regression)	Since the AUC for each class is 1, the classification is perfectly accurate.
Model Suitability Evaluation	Evaluate how well the model performs the classification task by comparing with the standard.	Suitable for the model, overall accuracy exceeds 90%.
Overfitting / Generalization Evaluation	Analyze whether the model is overfitted or generalized well enough by comparing training data and test data.	Not applicable for overfitting.
Implicit Evaluation	Evaluate how well the model balances precision and recall.	Balanced in terms of precision and recall.
Model Scalability Evaluation	Evaluate the scalability of the model by considering the expansion of decision boundaries and additional data.	Not applicable for scalability.
Interpretability	Whether the model is easy to interpret and provides clear decision trees or simple logic.	Random Forest offers the best interpretability due to visualized decision trees and detailed analysis.
Use Case Specificity	How well the model fits specific cases like classifying or predicting for a particular context.	Random Forest shows excellent performance in classifying and predicting specific events accurately.

신뢰성 있는 데이터를 기준으로 두집단의 정합성 결과가 동일하고, AI기반의 모델서능평가를 기반으로 주어진 데이터 셋에 한하여(양주시 미세먼지모니터링 시스템) 최적의 알고리즘을 Random Forest로 결론 도출 하였다.

연구 과정에서 AI 기반의 분류 모델 평기시 타당성과 유효성 해석에 부분을 표5로 정리되는 것을 확인하였다.

연구의 한계점으로는 연구 대상이 양주시로 한정되어 있어, 결과의 일반화 가능성이 제한적일 수 있다.

데이터의 시간적 범위가 명확히 제시되지 않아, 연구 결과의 시간적 유효성을 판단하기 어렵다.

외부 요인이 미세먼지 농도에 미치는 영향에 대한 변수가 양주시 데이터로 제한되어 모델의 장기적 성능 및 안정성에 대한 평가가 이루어지지 않았다.

REFERENCES

[1] Y. Kim, and K. Lee, "Accuracy Analysis of Machine Learning Methods for Predicting PM Concentration," J. Korean Soc. Atmos. Environ., Vol. 39, No. 2, pp. 149-164, April 2023. DOI: https://doi.org/10.5572/KOSAE.2023.39.2.149

[2] Z. Zhu, B. Chen, Y. Zhao, and Y. Ji, "Multi-sensing paradigm based urban air quality monitoring and hazardous gas source analyzing: a review," Journal of Safety Science and Resilience, Vol. 2, No. 3, pp. 131-145, September 2021. DOI: https://doi.org/10.1016/j.jnlssr.2021.08.004

[3] A. Kataria, and V. Puri, "AI- and IoT-based hybrid model for air quality prediction in a smart city with network assistance," IET Netw., Vol. 11, No. 6, pp. 221-233, February 2021. DOI: https://doi.org/10.1049/ntw2.12053

[4] Y. W. Lim, J. Eom, and K. Y. Kwahk, "Development of a water quality prediction model for mineral springs in the metropolitan area using machine learning," Korea Intelligent Information Systems Society, Vol. 29, No. 1, pp. 307-325, March 2023. DOI: 10.13088/jiis.2023.29.1.307

[5] S. G. Kim, "A Study on the AI-Based Fine Dust Prediction Model for Improving Reliability," Chonbuk National University, pp. 4-25, June 2023. DOI: 10.1234/abcd.5678

[6] J. E. Ware Jr, and B. Gandek, "Methods for testing data quality, scaling assumptions, and reliability: the IQOLA Project approach," Journal of Clinical Epidemiology, pp. 945-952, November 1998. DOI: 10.1016/S0895-4356(98)00085-7

[7] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, pp. 433-455, August 2006. DOI: 10.1007/978-0-387-45528-0

[8] C. A. U. Hassan, and M. S. Khan, "Comparison of Machine Learning Algorithms in Data classification," 2018 International Conference on Automation, Computational and Technology Management, pp. 945-952, April 2018. DOI: 10.23919/IConAC.2018.8748995

[9] A. A. Haruna, I. A. Mohammed, and A. Ahmad, "Machine learning predictive models for coronary artery disease," Journal of SN Computer Science, pp. 339-350, July 2021. DOI: 10.1007/s42979-021-00731-4

[10] E. Erdem, and F. Bozkurt, "A comparison of various supervised machine learning techniques for prostate cancer prediction," European Journal of Science and Technology, pp. 610-620, December 2021. DOI: 10.31590/ejosat.802810

[11] T. A. Alghamdi, and N. Javaid, "A survey of preprocessing methods used for analysis of big data originated from smart grids," Journal of IEEE Access, pp. 29149-29171, March 2022. DOI: 10.1109/ACCESS.2022.3156139

[12] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," Decision Analytics Journal, pp. 1-14, January 2023. DOI: 10.1016/j.dec.2023.1001819

Authors

Heungsup Sim received the B.S. Information and Communication Engineering, received the M.S.degrees in advertising science from Chung-Ang University, Korea, in 2010, M.S. and Ph.D. completion Dongyang University, Korea in 2013.

He is currently a Professor in the Computer & Military Department, Dongyang University. He is currently a professor of computer and military science at Dongyang University. He is interested in ESG data analysis, smart farm advancement, and IOT-cloud computing.