Journal of the Korea Academia-Industrial cooperation Society
Vol. 22, No. 7 pp. 403-408, 2021

https://doi.org/10.5762/KAIS.2021.22.7.403
ISSN 1975-4701 / eISSN 2288-4688

작물 생산량 예측을 위한 머신러닝 기법 활용 연구

김세원, 김영희*
호서대학교 융합공학과

A Study on the Application of Machine Learning Algorithm to Predict Crop Production

Se-Won Kim, Younghee Kim*
Division of Convergence Engineering, Hoseo University

요 약 기후변화로 인해 농작물 생산에 대한 관심이 증대되고 있고, ICT 기술을 접목한 스마트팜의 작물 생산량을 최적화하기 위한 노력이 증대되고 있다. 빅데이터를 활용한 작물과 환경 상태에 따른 생산 정도의 분석이 필요하고 나아가 생산량 예측을 위한 모델에 대한 연구가 필요하다. 3가지의 머신러닝 알고리즘 (Ridge Regression, Random Forest, XGBoost)을 후보 알고리즘으로 선정하여 작물 생산량 예측의 적합도를 평가 분석하였다. 실제값과 예측값의 오차를 산출한 MAE(Mean Absolute Error)/RMSE(Root Mean Square Error)값을 모델 평가 지표로 사용하여 알고리즘별 최적의 파라미터를 선정하였다. Ridge Regression의 파라미터 λ는 2.512이고, Random Forest의 파라미터는 분할 8, 트리 100이고, XGBoost의 파라미터는 감마 0, 깊이 10 이였다. 선정된 최적의 파라미터가 결정된 알고리즘들을 MAE/RMSE로 평가한 결과 XGBoost가 MAE 0.233, RMSE 0.817로 최소값을 나타내 최적 모델로 선정되었고, 변수 중요도를 확인해 본 결과, 재식밀도, 생장길이, 잎의 수 요인이 출하량 예측에 가장 중요한 요인으로 도출되었다. XGBoost 모델의 예측력을 R-Square 값을 통해 평가하였고, 각 작기별 총 출하량 예측 모델에 대하여 약 77%의 설명력을 보였다.

Abstract Climate change has increased interest in crop production and efforts to optimize crop production of smart farms incorporating ICT technology. Therefore, it is necessary to analyze the production level according to the crops and environmental conditions using big data. Hence, a model for predicting production is required. Three machine learning algorithms (Ridge Regression, Random Forest, and XGBoost) were selected as candidate algorithms to evaluate and analyze the fit of predicting crop production. The MAE (Mean Absolute Error) / RMSE (Root Mean Square Error) values, which calculated the error between the actual and predicted values, were used as model evaluation indices. The parameters with the smallest MAE/RMSE values were selected for the optimal model of each algorithm. Parameter λ of Ridge Regression was 2.512, the parameters of the Random Forest were division 8 and tree 100, and the parameters of XGBoost were gamma 0 and depth 10. An evaluation of the algorithms by MAE/RMSE revealed XGBoost to be the optimal model with minimum values of MAE = 0.233 and RMSE = 0.817. An examination of the importance of the variables revealed the planting density, growth length, and number of leaves to be important factors in forecasting shipments. The predictive power of the XGBoost model was evaluated using the R-Square value, which had an explanatory power of approximately 77% for the total shipment forecast model for each production period.

Keywords : Smart Farm, Prediction of Crop Production, XGBoost, Ridge Regression, Random Forest

*Corresponding Author : Younghee Kim(Hoseo Univ.)
email: younghee.km@gmail.com
Received May 13, 2021 Revised June 21, 2021
Accepted July 2, 2021 Published July 31, 2021

1. 서론

기후변화는 전 지구적인 기후가 변화하는 현상이지만, 최근 사회 문제로 떠오른 기후변화는 지구온난화를 가리킨다. 온실효과로 인한 지구 기온 상승 현상은 좁은 의미로는 인간 활동에 의해 19세기 말부터 지구의 평균 기온이 상승하는 현상을, 넓은 의미로는 지구의 기온이 어떠한 이유에서든 평균 이상으로 증가하는 현상을 뜻한다. 온실가스의 저감을 위한 탄소중립이 강조되면서 탄소배출을 줄이기 위한 도시 농업이 활성을 띄고 있다. 또한 농업의 기후지대가 변하기 시작하면서 농업생태계 전반이 달라지고 있다. 이러한 기후변화의 위험 속에서 농사 기술에 정보통신기술(ICT)을 접목하기 시작하였으며, 사물인터넷(IoT:Internet of Things) 기술을 이용하여 농작물 재배 시설의 온도, 습도, 햇볕량, 이산화탄소, 토양 등을 측정 분석하고, 분석 결과에 따라 제어 장치를 구동하여 적절한 상태로 변화시키는 지능화된 농장인 스마트팜을 통한 작물 생산이 증대되고 있다. 이러한 스마트팜의 많은 센서 장치들로부터 수집되는 다양한 데이터 중에서 어떤 인자들이 최종적으로 작물의 생산량에 많은 연관관계가 있는지를 살펴보고, 생산량을 예측할 수 있는 AI 기반의 최적 모델을 연구한다. 전체적인 연구의 흐름은 데이터 수집, 데이터 탐색, 분석 대상 품종 선정, 데이터 이상치 처리, 데이터 마트 구축, 분석 요인 선정(생산량과 연관 있는 요인), 예측 모델 후보 알고리즘 선정, 후보 알고리즘별 최적 파라미터 선정, 최적 예측 모델 선정, 예측 모델 변수 중요도 분석, 예측 모델 설명력 검증으로 진행하였다.

2. 본론

2.1 데이터 수집 및 마트 구축

농림수산식품교육문화정보원에서 제공하는 공공데이터 목록 34개를 확인하였으며, Table 1.과 같이 공공데이터 제공방식(FILE/API)에 따라 수급 가능한 목록 32개와 수급 불가능한 목록 2개를 식별하였다. 수급 가능한 목록 중 날짜 정보가 부재하거나, 비정형파일 형식이거나, 단순 목록 및 현황 정보 데이터이거나, 갱신이 중단되어 있는 데이터 21개는 분석 활용에 부적합한 목록으로 분류하였다. 개방되어 있는 공공데이터 목록 34개 중, 분석 활용에 이슈가 없는 목록 9개와 코드 정보 목록 3개를 분석 활용에 적합한 목록으로 선정하였다. 분석 활용에 이슈가 없는 9개의 데이터 목록 중 스마트팜 빅데이터가 포함되어 있으며, 해당 스마트팜 빅데이터는 5가지의 작기별 정보 서비스 데이터와, 10가지의 Provide 서비스 데이터로 구분되며, 데이터 특성에 따라 시간/일 단위로 제공되고 있다. 구체적인 내용은 Table 1.과 같다.

Table 1. Smart Farm Data List

	Data Name	Provided Information	Unit
Information Service Data	Crop Status Information	Date of Production Registered Sensor	day
	Environmental Information	Internal/External Information	time
	Control Information	Control Information for Each Sensor	time
	Growth Information	Growth Information by Item	day
	Management Information	Expenses Information	-
Provide Service Data	Identity Information	Item and Location Information	-
	Period Information of Cultivation	Corresponding to Farmhouse ID	day
	Environmental Information	Facility Gardening Environment Information	time
	Environmental Information (Pig)	Pig Raising Environmental Information	time
	Tomato/Paprika/Cucumber/Eggplant	Growth Information of Facility Horticultural Items (Used as an indicator for growth quality)	day
	Strawberry		
	Chrysanthemum		
	Oriental Melon		
	Pig	Growth Information of Pig Raising	day
	Management Information	Expenses Information	-

스마트팜 데이터의 작기 975개 중 가장 많은 작기별 정보를 제공하고 있는 품종인 토마토를 분석 대상 품종으로 선정하였다. 토마토의 163개 작기 정보 중 생육 정보를 측정하는 작기인 88개 작기 정보를 탐색하였으며, 해당 정보 중 환경정보를 측정하는 작기인 32개 작기 정보를 최종 분석 대상 데이터로 선정하였다. 해당 데이터의 Quality를 점검하였고, 농식품 백과사전을 기준으로 하여 범위를 벗어난 이상치 data는 기준 범위 내로 조정하였고, data가 '0'이거나, 누락되어 알 수 없는 항목은 제거하였다. 작기정보의 재식밀도가 누락되어 있는 경우, 작기정보의 '재배면적(㎡)'과 '식재된 총 작물의 수'를 활

용해 재식밀도 값을 생성하였다. (재식밀도 : 식재된 총 작물의 수(개) ÷ 재배면적(㎡)) 경영정보의 총 출하량이 누락되어 있는 경우 생육 정보의 '열매수'와 작기정보의 '식재된 총 작물의 수', '기준과중(g)'을 활용해 값을 생성하였다. (총 출하량(t) : 열매수(개)×기준과중(g)×식재된 총 작물의 수(개)) '열매수'는 작기별 생육정보 마지막 측정일의 표본 4개 평균값을 활용하였다. 생육정보와 환경정보가 중첩되는 기간이 3개월로 분석되어 분석 기준 기간을 3개월로 산정하였고, 생육정보는 월 단위로 변환하였다. 이렇게 변환된 데이터의 작기별 row수는 분석에 부족하여 Random Over-Sampling을 통한 데이터를 확장하였다.

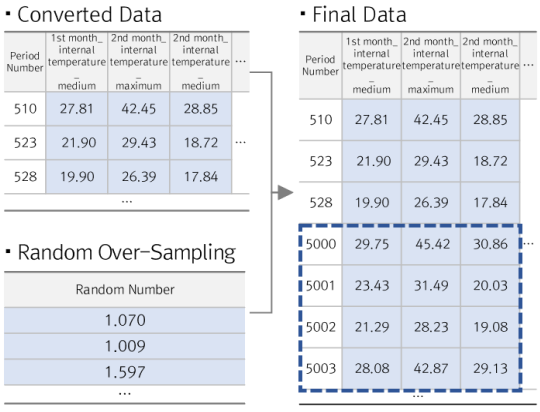
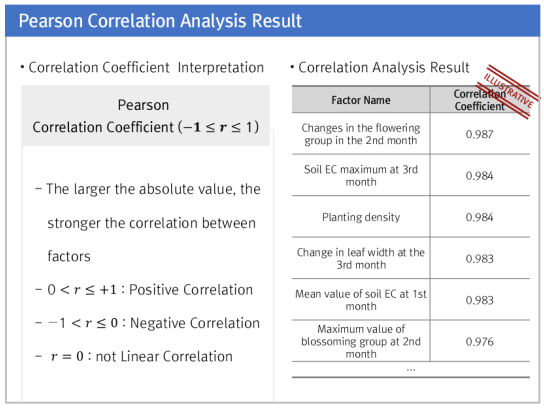


Fig. 1. Data Conversion

Fig. 1.과 같이 이상치를 처리하고 일부 항목의 데이터 보완을 통한 데이터 전처리를 진행하였으며, 전처리된 데이터의 분석 기간 선정, 데이터 변환, 데이터 확장을 통해 207개 항목을 가진 최종 데이터 마트를 구축하였다.

2.2 요인 선정 방법론 및 모델 적용 최종 요인 선정 결과

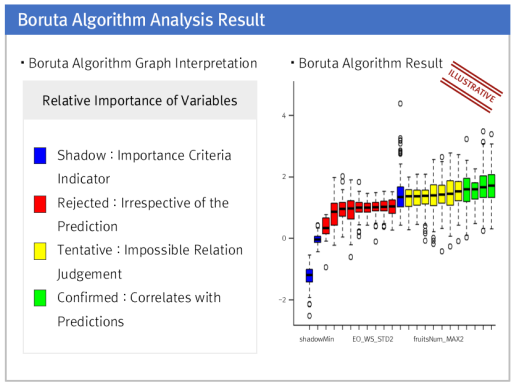
최적 예측 모형 구축을 위해 총 출하량에 영향을 미치는 요인을 선정하는 것이 필요하며, 요인 선정 방법으로 상관분석과 Boruta 알고리즘을 활용하였다. 먼저 통계 기반 요인 선정 방법 중 하나로 연속형 요인 간 선형 관계를 확인할 수 있는 Pearson 상관분석을 수행하였고, Pearson 상관계수의 절대값이 0.7 이상인 강한 상관관계를 가지는 요인을 Fig. 2.와 같이 136개 추출하였다.



Factor Name	Correlation Coefficient
Planting Density	0.98
Average value of growth length at 1st month	-0.77
Maximum stem diameter at 1st month	-0.74
Maximum value of internal humidity at 2nd month	0.25
...	

Fig. 2. The Result of the Analysis by Pearson Correlation

머신러닝 기반의 요인 선정 알고리즘 중 하나인 Boruta 알고리즘을 사용하였으며, Boruta 알고리즘은 변수의 'shadow 속성'의 최대 Z-score 보다 높은 Z-score 가 존재할 시 'Confirmed'로 판단하였으며, 예측과 상관있는 유의미한 요인을 Fig. 3.과 같이 90개 추출하였다.



Factor Name	Boruta Result
Planting Density	Confirmed
Average value of growth length in 1st month	Confirmed
Maximum stem diameter at 1st month	Tentative
Maximum value of internal humidity at 2nd month	Rejected
...	

Fig. 3. The Result of the Analysis by Boruta Algorithm

위의 두 과정에서 추출된 Pearson 상관분석 결과에 따른 136개의 요인과, Boruta 알고리즘 활용 결과에 따라 도출된 90개 요인간의 공통된 요인을 Fig. 4.와 같이 도출하였다.

Factor Name	Correlation Coefficient	Boruta Result	Selection Result
Planting Density	0.98	Confirmed	O
Average value of growth length in 1st month	-0.77	Confirmed	O
Average value of leaf length at 1st month	0.24	Confirmed	X
Maximum stem diameter at 1st month	-0.74	Tentative	X
Maximum value of internal humidity at 2nd month	0.25	Rejected	X
Average number of leaves at 3rd month	-0.77	Tentative	X
Average value of number of fruits at 3rd month	-0.56	Rejected	X
Change in height of flower cluster in 3rd month	-0.94	Rejected	X
...			

Fig. 4. The Selection Result of Common Factor

그 결과, 예측 모델에 적용할, 총 출하량에 영향을 미치는 63개(30.29%)의 요인을 최종 선정하였으며, 재식밀도, 1개월차 생장길이 평균값, 2개월차 잎 수 평균값이 가장 중요한 3가지 요인으로 도출되었다.

2.3 최적화 모델선정

작물 출하량 예측에 대한 문헌연구 및 데이터 특성을 고려하였고, 생산량과 연관된 환경 요인 값의 관계가 중요하여 종속변수와 독립변수의 상관관계를 모델링 하는 회귀분석인 Ridge Regression 알고리즘을 첫 번째 후보 알고리즘으로 선정하였고, test 단계에서 나온 결과와 ground truth를 비교하여 accuracy를 측정하는, 분류와 회귀 모두 가능한 지도 학습 모델 중 하나인 Random Forest를 두 번째 후보 알고리즘으로 선정하였고, 선형 알고리즘과 Tree 알고리즘을 모두 사용하여 더욱 강력한 알고리즘을 만들 수 있는 XGBoost 알고리즘을 세 번째 후보 알고리즘으로 선정하였다.

2.3.1 모델별 최적 수식 도출

각 모델별 Validation Set을 활용한 최적화 모델 파라미터를 선정하였다. 모델 평가 방법으로 분류 문제를 평가할 수 있는 Confusion Matrix, Accuracy, Precision, Recall, F1-Score, AUC 등이 있으나, 회귀 문제를 평가하는 방법이 타당하다고 판단하여, MAE(Mean Absolute Error)와 RMSE(Root Mean Square Error) 값을 모델 최적화 평가 지표로 사용하였다. MAE는 모든 절대 오차의 평균이므로 실제 출하량과 예측 출하량 값의 차이의 절대값 오차를 확인하는 것이며, 수식은 다음과 같다.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \quad (1)$$

MAE와 더불어 인공지능 예측 모델의 정확도를 측정하기 위해 사용되는 대표적인 지표 중 하나로서 인공지능 예측의 오차를 하나의 숫자로 표현해줌으로써 예측 모델의 정확도를 이해하기 쉬운 RMSE는 예측값과 실제값의 차이를 구한 후 모든 오차들의 제곱의 평균값을 구하고 평균값의 제곱근을 구하면 RMSE 값을 찾을 수 있으며, 수식은 다음과 같다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - acutal_i)^2}{n}} \quad (2)$$

각 알고리즘별 모델 최적화 파라미터 조정 학습 결과는 Table 2.와 같으며 Ridge Regression은 람다 2.512에서 MAE와 RMSE가 최솟값을 보였고, Random Forest는 분할 8과 트리 100에서 MAE와 RMSE가 최솟값을 보였으며, XGBoost는 감마 0와 깊이 10에서 MAE와 RMSE가 최솟값을 보여 최적 파라미터로 선정되었다.

Table 2. The Evaluation Result by MAE/RMSE

Algorithm	Parameter	MAE	RMSE
Ridge Regression	lambda 79.433	5.106	6.493
	lambda 70.424	5.002	6.381
	lambda 37.614	4.889	5.902
	lambda 11.182	4.475	5.836
	lambda 2.512	4.445	5.762
Random Forest	division 8, tree 500	0.905	0.225
	division 8, tree 100	0.694	0.196
	division 10, tree 500	0.554	1.152
	division 6, tree 300	0.845	0.525
	division 4, tree 200	0.460	1.164
XGBoost	gamma 0, depth 6	0.518	0.199
	gamma 0, depth 10	0.507	0.198
	gamma 0, depth 50	0.733	0.279
	gamma 0.1, depth 10	0.522	0.281
	gamma 0.25, depth 30	0.702	0.256

2.3.2 최적 모델선정

선정된 최적화 파라미터 적용 모델 성능을 Test Set을 활용하여 검증해 보았고, MAE/RMSE를 적용하여 평가한 결과 XGBoost가 최솟값을 가져 성능이 가장 우수하게 나타났다.

Classification	Parameter	MAE	RMSE
Ridge Regression	lambda 2.512	3.815	3.815
Random Forest	division 8, tree100	0.308	0.876
SELECT XGBoost	gamma 0, depth 10	0.233	0.817

Fig. 5. The Evaluation Result for Best Estimate Model

2.4 예측 모델 검증

R-Square 값을 통한 모델의 설명력 확인 결과, 전체 변수를 활용한 모델의 경우 78.39%의 설명력을 가지며, 상위 10%의 중요 요인을 활용한 모델의 경우 75.95%의 설명력을 보였다. 전체 31개의 요인을 활용한 경우와 3개의 요인을 활용한 경우의 설명력 차이가 크지 않으므로 모델 해석 및 다중공선성 해결을 위해 상위 10%의 요인을 활용한 모델이 적합하다고 판단된다. 위 두 가지 결과를 통해 각 작기별 총 출하량 예측 모델에 대하여 약 77%의 설명력을 보인다.

3. 결론

본 연구를 통하여 분석에 활용할 중요 요인을 분석한 결과 최종 선정된 63개 요인 중 재식밀도, 1개월차 생장길이 평균값, 2개월차 잎수 평균값 3가지 생육정보가 가장 중요한 요인임을 알 수 있었다. 63개 요인을 적용하여 학습한 모델을 평가한 결과, 3가지 알고리즘의 최적 파라미터를 평가한 결과, Ridge Regression은 람다 2.512, Random Forest는 분할8 트리100, XGBoost는 감마0 깊이10이 모델별 최적의 파라미터로 선정되었다. 해당 파라미터를 적용한 3개 모델을 평가한 결과 XGBoost가 가장 성능이 좋은 것으로 나타났고, R-Square를 통해 확인한 설명력은 약 77%였다. 따라서 이러한 AI 학습 모델을 적용한 알고리즘을 통해 농작물 생산 예측이 가능한 것으로 보인다. 생육 정보 데이터 품질 관리를 위한 표준화된 가이드가 필요하다는 것도 알 수 있었다. 다른 작물에 대한 예측을 위해 작물별로 주요 요인에 대한 공통된 IoT 정보의 수집이 필요하다고 판단되었다. 또한 병해충 정보와 같이 농작물에 영향을 미치는 추가적인 데이터가 수급된다면 예측 모델이 개선될 수 있을 것이라 생각된다. 이러한 예측 모델을 통해 스마트팜에서 생산되는 작물의 품질 향상과 출하량이 증대되는데 기여함으로써 기후 변화로 인한 작물 생산 저하 이슈를 해결하는 데 있어 일조할 것을 기대하며, 작물 생산 시설을 소비자와 가까운 거리에 설치 가능한 스마트팜을 통해 유통 거리 감소의 효과로 탄소배출도 줄일 수 있는 가능성을 기대한다. 향후 타 품종에 관한 연구, 경영정보를 활용한 경제성분석, 탄소중립에 기여할 수 있는 모델을 연구하겠다.

References

[1] Kursa M., Rudnicki W., "Feature Selection with the Boruta Package" Journal of Statistical Software, Vol. 36, Issue 11, Sep 2010

[2] Saptashwa Bhattacharyya, Ridge and Lasso Regression: L1 and L2 Regularization, Sep 26, 2018

[3] Baek Kyun Shin, Machine Learning-Decision Tree, Jul 2019

[4] Dev Cristoval, Random Forest Algorithm, Jan 2019

[5] Rokach, L.; Maimon, O., "Top-down induction of decision trees classifiers-a survey", 2005

[6] Murthy S., Automatic construction of decision trees from data, 1998

[7] Deng,H.; Runger, G.; Tuv, E. (2011). 《Bias of importance measures for multi-valued attributes and solutions》. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). P293;300.

[8] Papagelis A., Kalles D.(2001). Breeding Decision Trees Using Evolutionary Techniques, Proceedings of the Eighteenth International Conference on Machine Learning, p.393-400, June 28-July 01, 2001

[9] Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "Bayesian CART model search." Journal of the American Statistical Association 93.443 (1998)

[10] Jamie Shotton (January 2013). "Real-time human pose recognition in parts from single depth images". 《CACM》

[11] Antonio Criminisi (September 2010). "Regression Forests for Efficient Anatomy Detection and Localization in CT Studies".

[12] Darko Zikic (October 2012). "Decision Forests for Tissue-Specific Segmentation of High-Grade Gliomas in Multi-channel MR". 《MICCAI》

[13] Datacamp tutorial

[14] Vishal Morde, XGBoost Algorithm, Apr 8 2019

[15] Analytics Vidhya

[16] IPCC Special Report on the Impacts of Global Warming of 1.5 °C

[17] United Nations Framework Convention on Climate Change

[18] Korea Environment Corporation, Climate Change Public Relations Portal

[19] NASA: Climate Change and Global Warming

[20] Climate change – Wikipedia

[21] IPCC — Intergovernmental Panel on Climate Change

[22] Nature Climate Change

[23] A new paradigm for climate change K Anderson, A Bows - Nature Climate Change, 2012

[24] Predictive algorithms and AI

[25] TO Ayodele; Types of machine learning algorithms - New advances in machine learning, 2010

[26] Jenna Burrell; How the machine 'thinks': Understanding opacity in machine learning algorithms

[27] Ayon Dey; Machine Learning Algorithms: A Review

[28] Mohssen Mohammed, Muhammad Badruddin Khan, Eihab Bashier Mohhamed Bashier; Machine Learning Algorithms and Applications

[29] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O., Akinjobi J.; Supervised Machine Learning Algorithms: Classification and Comparison

[30] Susmita Ray; A Quick Review of Machine Learning Algorithms

[31] Ministry of Agriculture, Food and Rural Affairs, Republic of Korea; Smart Farm Dispersion Method

[32] Smart Farming Technology - Advanced Agriculture Solution;

[33] Korea University Business School(2013). Big Data Use Case and Implications

[34] Korea Policy Briefing; Smart Farm

[35] Inha University; Trends and Prospects of the Climate Change Convention, Global Warming Trend

[36] Meteorological Agency; Climate Information Portal

김 세 원(Se-Won Kim) [정회원]

- 2001년 2월 : 성균관대학교 공과대학 기계설계학과 (공학석사)
- 2008년 2월 : 고려대학교 경영전문대학원 (경영학석사)
- 2001년 1월 ~ 2012년 4월 : 삼성 SDS 책임
- 2012년 4월 ~ 2020년 10월 : SAP Korea 상무
- 2020년 10월 ~ 현재 : 솔트룩스 상무

〈관심분야〉
빅데이터, 인공지능, 클라우드, SaaS

김 영 희(Younghee Kim) [정회원]

- 2004년 8월 : 서울대학교 보건대학원 (환경보건학석사)
- 2008년 8월 : 서울대학교 보건대학원 (환경보건학박사)
- 2013년 3월 ~ 현재 : 호서대학교 벤처대학원 부교수

〈관심분야〉
대기환경, 수질환경, 환경에너지