Journal of KIIT. Vol. 20, No. 12, pp. 49-55, Dec. 31, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.12.49

약용작물의 유효성분 함량 예측을 위한 기계학습 모형의 최적화 및 요인 분석

허 진 경*

Optimization of Machine Learning Model and Factor Analysis for Predicting the Active Ingredient Content of Medicinal Crops

JinKyoung Heo*

본 성과물은 산업통상자원부-한국산업기술진흥원-지역혁신클러스터육성(R&D)-국가혁신융복합단지지원 연구사업(과제번호: P0015318)의 지원으로 이루어진 것임

요 약

식물의 생물학적 화학 성분은 합성 화학 성분보다면 부작용과 내약품성이 적다. 더욱이 생약기술의 발전으로 약용작물이 주목받고 있다. 그리고 약용작물의 집약적 재배가 가능하고 환금성이 기대되므로 농업 경제에 주는 영향이 점점 더 높아지고 있다. 따라서 이 논문은 기계학습 모형의 최적화를 통해 생육환경이 유효성분의 함량 증대에 영향을 줄 수 있음을 확인하고 요인 분석을 통해 최적의 생육환경을 찾는 방법을 제안한다. 이 논문은 실험을 위한 대상 작물로 밀싹을 선택했다. 생육환경이 밀싹의 유효성분인 이소오리엔틴에 영향을 주는지 확인하기 위해서 기계학습 모형의 최적화를 실시했고, 이소오리엔틴의 함량을 증가시킬 수 있는 재배환경을 찾기 위한 요인 분석 방법을 제안한다.

Abstract

Biochemical components of plants have less side effects and chemical resistance than synthetic chemical components. Moreover, medicinal crops are attracting attention due to the development of herbal medicine technology. In addition, since intensive cultivation of medicinal crops is possible and cash exchangeability is expected, the impact on the agricultural economy is increasing. Therefore, in this paper, we confirm that the growth environment can affect the increase in active ingredient content through optimization of machine learning models, and propose a method to find the optimal growth environment through factor analysis. This paper selected wheatgrass as the target crop for the experiment. To determine whether the environment affects isoorientin, machine learning model was optimized, and factor analysis method is proposed to find a environment that can increase the content of isoorientin.

Keywords

machine learning, feature importance, factor analysis, isoorientin, active ingredient content prediction

* 연세대학교 산학협력단(공학계열) 연구교수
- ORCID: https://orcid.org/0000-0001-7279-0994

· Received: Oct. 31, 2022, Revised: Nov. 18, 2022, Accepted: Nov. 21, 2022
· Corresponding Author: JinKyoung Heo
Dept. of University Industry Foundation, Yonsei University, Korea
Tel.: +82-2-2123-2147, Email: hjk7902@yonsei.ac.kr

Ⅰ. 서 론

작물 수확량을 예측하는 것은 물, 자외선(UV), 살충제, 비료 및 해당 지역에 적용되는 토지 면적과 같은 많은 매개변수에 따라 달라지기 때문에 쉬운 일이 아니다[1].

식물은 그 자체뿐만 아니라 다른 유기체에서도 생물학적으로 중요한 많은 화학 물질을 생산한다. 이러한 화학 물질 중 일부는 자신의 생존을 향상하게 시키며, 그 과정에서 화학 물질이 더 많이 생산될 수 있는데, 식물이 어떻게 하면 같은 생육 기간에 더 많은 화학 물질을 포함할 수 있을지에 관한 연구가 진행되고 있다[2].

이 논문은 약용 식물의 유효성분 함량 예측 및 요인 분석에 기계학습 모형이 사용될 수 있음을 알아본다. 실험에 사용한 작물은 밀싹으로 정했다. 밀싹은 주로 영양소의 농축된 공급원으로 사용된다. 그것은 비타민 A, 비타민 C, 그리고 비타민 E, 철분, 칼슘, 마그네슘, 그리고 아미노산을 풍부하게 함유하고 있다. 그리고 이소오리엔틴과 플라본배당체 등의 폴리페놀류가 100g당 최대 약 1,360mg 들어있다. 밀싹은 단시간에 길러 수확할 수 있어 농가소득 증대에 이바지할 수 있는 작물이다[3].

이 논문의 구성은 다음과 같다. 2장에서는 기계학습과 이를 기반으로 변수 탐색 및 요인 분석과 관련된 연구를 분석한다. 3장에서는 기계학습에 사용할 데이터의 수집 및 전처리와 실험군과 대조군의 통계량을 비교한다. 4장에서는 기계학습 모형을 만들고, 요인을 분석하여 이소오리엔틴에 영향을 주는 변수의 구간을 찾는다. 마지막으로 5장에서는 결론 및 향후 연구를 서술한다.

Ⅱ. 관련 연구

2.1 기계학습과 앙상블

[4]는 작물의 수확량을 분석하기 위해 두 가지 다른 기계학습 알고리즘을 사용했다. SVR(Support Vector Regression) 및 LR(Linear Regression)이라는 두 가지 알고리즘은 연속형 변수의 추정을 예측하기 위한 회귀 모형이고, 이 회귀 모형의 평가를 위해 평균 제곱 오차(MSE) 및 결정 계수(R2)를 사용했다.

기계학습의 앙상블은 기계학습 연구의 주요 현재 방향 중 하나를 구성하며 광범위한 실제 문제에 적용되었다. 앙상블에 대한 통일된 이론이 없음에도 불구하고 여러 기계학습 모형을 결합하는 데에는 많은 이론적 이유가 있으며 이 접근법의 효과에 대한 경험적 증거가 있다. 그리고 앙상블 모형을 사용하면 단일알고리즘인 SVR과 LR 알고리즘들보다 더 나은 일반화된 결과를 얻을 수 있다[5].

그래서 본 논문에서는 앙상블 모형을 이용해서 밀싹의 생육 환경에 따른 이소오리엔틴 함량 예측을 위한 회귀 모형을 만들었으며, 모형의 평가를 위해 [4]에서 사용한 결정계수(R2)를 적용했다.

2.2 변수 탐색 및 요인 분석

Desboulets Loann은 단순한 선형 구조에서 복잡한 비모수 구조 등 여러 모형에 대해 시험 기반, 패널티 기반, 선별 기반으로 분류하여 변수 탐색을 진행하였다. 그리고 연구에서 모형의 형태와 데이터 특수성(공통성, 그룹 등) 및 목표에 의존해야 함을 보였다. 그리고 경험적 연구에서 널리 사용되는 방법에 대한 선택 일관성이 논의되었고 몇 가지 개선사항이 제시되었다. 많은 데이터를 사용할 수 있게 된 지금 모델 선택 영역은 여전히 연구되고 있다. 그런데도, 많은 수의 변수를 처리하는 방법은 모델 복잡성 측면에서 제한된다. 이는 주로 차원의 저주 때문이며 고차원에서 매우 복잡한 모델을 찾는 것을 방지한다[6].

특징 변수들을 찾고 그 특징 변수들의 중요도를 측정하는 방법에는 여러 가지 다른 접근 방식이 있다. Saarela와 Jauhiainen은 로지스틱회귀 및 랜덤포레스트 방법을 사용하여 변수 중요도 측정값을 비교했다. 그 결과 가장 중요한 변수는 사용하는 알고리즘 따라 다르다는 것을 보여준다. 그리고 몇 가지 설명 기술의 조합이 더 신뢰할 수 있는 결과를 제공할 수 있다고 주장했다[7].

이 논문에서는 XGBoost, LightGBM 그리고 랜덤포레스트 기계학습 회귀 모형을 이용해서 조도, 온도, 습도, 이산화탄소가 밀싹의 유효성분인 이소오리엔틴 함량과의 연관성을 확인하였고, 의사결정나

무 모형을 시각화하여, 각 변수가 어떤 값을 가졌을 때 이소오리엔틴 함량이 가장 많은지를 찾았다.

III. 데이터

3.1 데이터 수집

밀 종자 금강, 백강, 앉은뱅이, 새금강, 조경 중 이소오리엔틴 함량이 가장 높은 금강 종자를 시나리오에 맞게 통제된 환경에서 재배하였다. 재배 환경에서 다르게 제공하는 것은 빛의 색, 물의 산성도, 빛의 밝기, 온도, 습도, 이산화탄소 등이 있다.

작물의 성장에 가장 큰 영향을 줄 수 있는 것이 빛과 물이므로 빛의 색과 물의 산성도에 따라 대조군과 실험군으로 나눠서 그 차이를 먼저 확인해야 했다. 빛의 색에 따라 백색광과 청색광 환경에서 자라는 밀싹 데이터를 대조군과 실험군으로 나눠 수집하였고, 물의 산성도에 따라 ph 7.3인 수돗물과 ph 6.9인 세척수에서 자라는 밀싹 데이터를 대조군과 실험군으로 나눠 수집하였다.

일관성 있고 외부 영향을 받지 않는 데이터를 확보하기 위해서 그림 1의 밀싹 전용 재배기를 사용하였으며, 데이터 수집 기간은 2021년 3월 1일부터 2021년 7월 22일까지이며, 센서 데이터 총 약 98만 건 데이터를 수집했다.

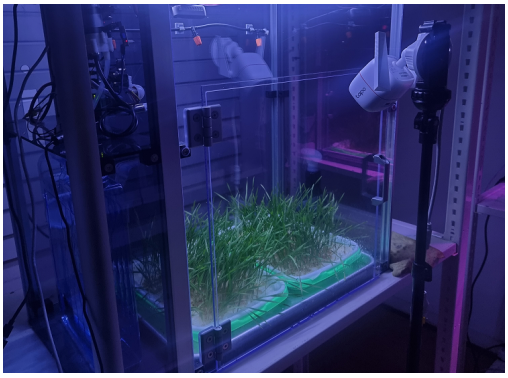

그림 1. 밀싹 재배기
Fig. 1. Wheatgrass incubator

대조군과 실험군 모두 센서 데이터는 식물재배기의 고유번호, 날짜, 온도, 습도, 이산화탄소 농도, 조도 데이터를 가지고 있다. 데이터에는 측정된 이소오리엔틴 함량 변수가 종속변수로 설정되어 있다.

그림 2의 시나리오 데이터의 주요 변수에는 scenario(실험군/대조군), water(물의 산성도), led(빛의 색과 조도), ingredient(최종 수확물의 그램 당 이소오리엔틴 함량(mg)) 등이 있고, 센서 데이터는 IDX(인덱스), S_N(식물재배기의 고유번호), air_temp(온도), air_humid(습도), air_co2(이산화탄소 농도), air_lux(조도), DT(날짜) 정보를 포함하고 있음을 보여주고 있다.

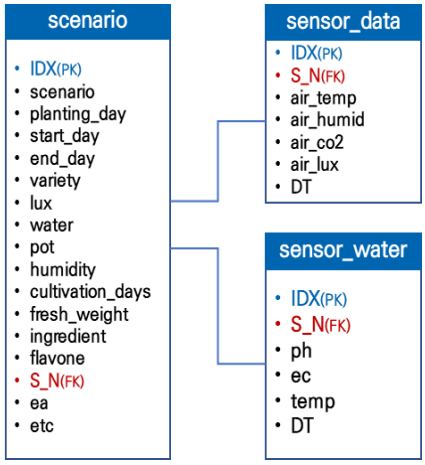

그림 2. 데이터 구조
Fig. 2. Data structure

3.2 데이터 전처리와 기초통계량

그림 3은 통계량에서 유의미하지 않은 인덱스, 재배기번호, 날짜 데이터는 제외하였고, 대조군과 실험군 별로 이소오리엔틴 함량과 물의 산성도(ph) 정보가 결측치 데이터를 분석에서 제외한 결과이다.

	IDX	S_N	air_temp	air_humid	air_co2	air_lux	DT	ingredient	ph
452192	595775	2F_159	29.3304	57.942001	492	3	2021-07-20 09:19:00	1.36	6.9
452194	595777	2F_172	28.483101	52.174099	593	3	2021-07-20 09:19:00	0.89	6.9
452199	595782	2F_157	28.601101	60.101101	444	6	2021-07-20 09:19:00	1.02	6.9
452206	595789	2F_171	29.105101	57.835098	528	4	2021-07-20 09:19:00	0.89	6.9
452208	595791	2F_160	28.740499	61.726101	449	2	2021-07-20 09:19:00	1.36	6.9
...	...	...	...	...	...	...	...	...	...

그림 3. 데이터 라벨링 및 전처리
Fig. 3. Data labeling and preprocessing

각 그룹 내에서 차이가 없는 물의 산성도(ph)를 제외하고 대조군과 실험군 센서 데이터들의 기초통계량을 확인하였다.

표 1과 표 2는 온도, 습도, 이산화탄소, 조도 등의 평균, 표준편차, 사분위수를 보여준다.

표 1. 대조군의 기초통계량
Table 1. Basic statistics of comparative group

	mean	std	25%	median	75%
temperature	26.76	1.79	25.54	26.10	27.79
humidity	60.63	5.10	57.30	60.44	63.67
CO2	472.34	76.07	409.00	450.00	507.00
brightness	12.37	29.19	0.00	0.00	2.00

표 2. 실험군의 기초통계량
Table 2. Basic statistics of experimental group

	mean	std	25%	median	75%
temperature	26.91	1.97	25.68	26.26	28.12
humidity	60.04	5.36	56.40	59.88	63.55
CO2	472.16	148.89	469.00	469.00	544.00
brightness	108.19	306.32	1.00	1.00	3.00

그림 4는 대조군과 실험군에서 변수들의 상관계수를 통해 다중공선성이 없음을 확인한다.

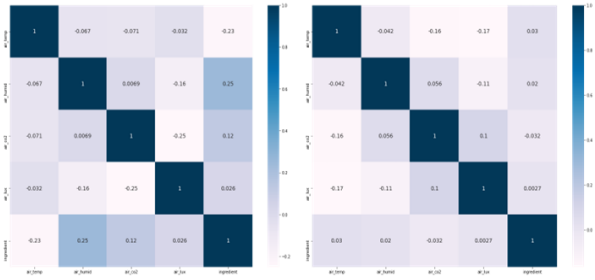

그림 4. 대조군(좌) 및 실험군(우)의 상관계수 히트맵
Fig. 4. Corr. heatmap of comparative group(left) and experimental group

3.3 대조군과 실험군

대립가설을 '대조군과 실험군의 이소오리엔틴 함량에는 유의미한 차이가 있다'라로 정의하며 귀무가설은 '대조군과 실험군의 이소오리엔틴 함량에는 유의미한 차이가 없다'라고 정의했으며, 그림 6의 t-test를 통해 검증했을 경우 p-value는 0.0으로 귀무가설을 기각하므로 대조군과 실험군 사이의 이소오리엔틴 함량 변화는 차이가 없는 것으로 나타났다. 이것은 빛의 색과 물의 산성도가 다른 환경에서 수집한 실험 데이터를 이용해서 이소오리엔틴 함량을 예측하는 모형을 만들어도 유의미한 결과를 얻을 수 있음을 의미한다.

```
from scipy import stats

t_test1 = stats.ttest_ind(con['ingredient'], expe['ingredient'], equal_var=False)
print(t_test1)
```

```
Ttest_indResult(statistic=346.00064659046916, pvalue=0.0)
```

그림 5. t-test 통계량과 p-value
Fig. 5. t-test statistics and p-value

Ⅳ. 기계학습 및 요인 분석

4.1 모형에 사용한 알고리즘

기계학습 모형은 앙상블을 사용했으며, 앙상블 모형 중에 부스팅 알고리즘에 해당하는 XGBoost와 LightGBM 그리고 배깅 알고리즘에 해당하는 랜덤포레스트를 사용했다. 모형의 평가는 회귀모형을 사용하므로 결정계수(R2)를 사용하였다.

XGBoost는 부스팅 기반 알고리즘인데, 트리를 이용한 부스팅은 매우 효과적이고 널리 사용되는 기계학습 방법이다. XGBoost는 확장 가능한 종단간 트리 부스팅 알고리즘이다. 이것은 데이터 과학자들이 많은 기계학습 문제에 대한 최신 결과를 달성하기 위해 널리 사용한다. XGBoost는 희소 데이터에 대한 새로운 희소성 인식 알고리즘과 근사 트리 학습을 위한 가중 분위수 스케치를 사용한다[8].

XGBoost 모형은 매개변수가 booster: gbtree, eta: 0.2, max_depth: 15, subsample: 0.8일 경우에 최적화되었고 이때 결정계수가 0.868이다.

XGBoost에 사용한 booster 매개변수의 값인 bgtree는 GBDT(Gradient Boosting Decision Tree)를 의미한다. GBDT는 차원이 높고 데이터 크기가 큰 경우 모든 데이터 객체를 스캔하여 가능한 모든 분할 지점의 정보 이득을 추정해야 하므로 시간이 오래 걸리므로 효율성과 확장성이 떨어진다. 그런데 LightGBM은 GOSS(Gradient-based One-Side Sampling) 및 EFB(Exclusive Feature Bundling)를 사용하여 구현한 GBDT이다.

GOSS를 사용하면 기울기가 작은 데이터 객체의 상당 부분을 제외하고 나머지만 정보 이득을 추정하는 데 사용한다. 더 큰 기울기를 가진 데이터 객체가 정보 이득 계산에서 더 중요한 역할을 하므로 GOSS는 훨씬 더 작은 데이터 크기로 정보 이득의 매우 정확한 추정을 얻을 수 있다. EFB를 사용하면 상호 배타적인 기능을 번들로 묶어 기능의 수를 줄인다[9].

LightGBM을 이용한 모형은 max_depth가 16이며 나머지 파라미터는 기본값일 경우 최적화되었고, 이때 결정계수가 0.752이다.

의사결정나무는 높은 실행 속도로 분류하지만, 전통적인 방법으로 파생된 트리는 종종 보이지 않는 데이터에 대한 일반화 정확도의 손실 가능성에 대해 임의의 복잡성으로 성장할 수 없다. 복잡성에 대한 제한은 일반적으로 훈련 데이터에 대한 차선의 정확도를 의미한다.

랜덤포레스트는 확률적 모델링 원칙에 따라 훈련 데이터와 보이지 않는 데이터 모두에 대한 정확도를 높이기 위해 용량을 임의로 확장할 수 있는 의사결정나무의 앙상블 구축을 기반으로 하는 비선형 분류 및 회귀를 위한 알고리즘이다. 랜덤포레스트는 특징 공간의 무작위로 선택된 부분 공간에 여러 트리를 구축하는 것이다. 서로 다른 부분 공간에 있는 나무는 보완적인 방식으로 분류를 일반화하고 결합된 분류는 단조롭게 개선될 수 있다[10][11].

랜덤포레스트는 기본 모형이 최적화된 결과를 도출하였고, 이때 결정계수가 0.876으로 XGBoost, LightGBM 보다 더 높은 평가 점수를 얻었다.

4.2 요인 분석

배깅 알고리즘인 RandomForest의 결정계수는 0.876, 부스팅 알고리즘 XGBoost의 결정계수는 0.868, LightGBM 알고리즘의 결정계수는 0.752였다. 이를 통해 우리는 기계학습 모형을 통해 조도, 온도, 습도, 이산화탄소가 이소오리엔틴 함량에 영향을 준다는 것을 확인했다.

XGBoost의 booster수가 linear일 때는 결정계수가 0.17로 매우 낮게 나왔고, gbtree일 경우 모형이 0.876으로 더 높게 나왔으며, RandomForest는 기본적으로 트리 모형을 기본 모형으로 사용하고 있으므로 회귀모형의 변수별 영향도와 종속변수에 영향을 주는 구간을 찾기 위해 대표적인 트리 모형인 의사결정나무 모형을 이용해서 이소오리엔틴 함량에 영향을 주는 요인을 탐색하였다.

그림 6은 센서 데이터변수 4개의 구간을 탐색하기 위해서 깊이를 5로 설정한 트리이며, 그 결과 조도는 87.5보다 크고, 온도는 26.183보다 높고, 습도는 53.338보다 크며, 이산화탄소 농도는 412.5보다 높을 때 이소오리엔틴 함량의 평균은 1.35로 가장 높았다. 이때의 샘플의 수는 1,426개이고, 제곱오차는 0.004였다.

그림 7은 깊이를 4로 설정한 트리이고, 이 트리는 이산화탄소 함량 변수가 제거되고 이때 이소오리엔틴 함량은 1.332이며 샘플의 수는 2,354개, 제곱오차는 0.012였다.

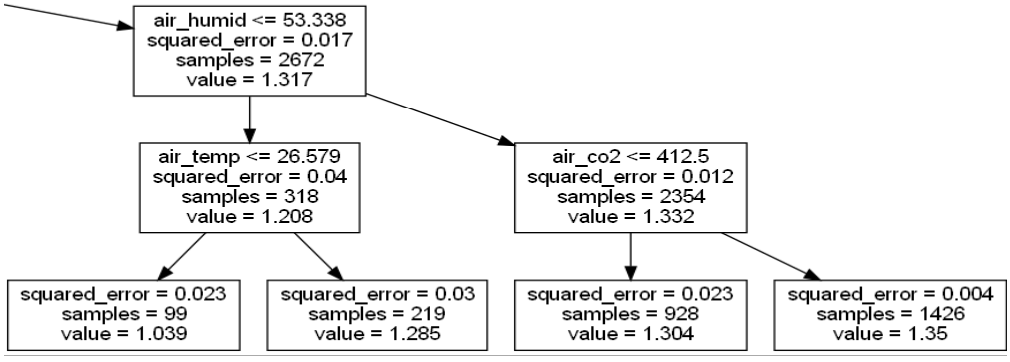

그림 6. 깊이 5인 트리의 오른쪽 끝
Fig. 6. Right leaf node of tree with depth 5

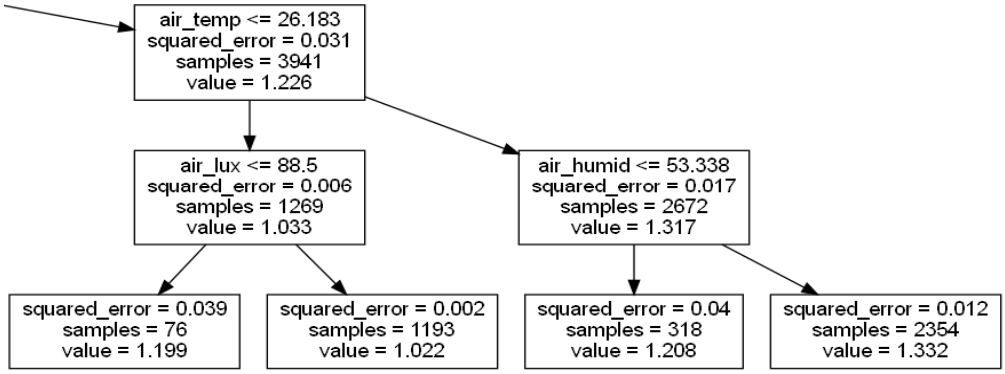

그림 7. 깊이 4인 분류 트리의 오른쪽 끝
Fig. 7. Right leaf node of tree with depth 4

V. 결론 및 향후 연구

본 연구는 기계학습 모형의 최적화를 통해 밀싹의 생육 환경이 유효성분 함량에 영향을 준다는 것을 확인했으며, 밀싹의 이소오리엔틴 함량을 높일 수 있는 요인들을 찾고 해당 변수들의 구간을 찾는다. 이 논문은 기계학습 모형의 독립변수와 목표변수 간의 인과관계를 통해 중요 요인을 찾는 것이 아닌 의사결정나무 모형 시각화를 통해 이소오리엔틴 함량에 영향을 주는 요인과 그 범위를 찾을 수 있었다.

그 결과 2진 분류트리의 깊이를 5와 4로 이소오리엔틴에 영향을 주는 요인들의 범위를 알아본 결과를 정리하면 다음 표 3과 같다.

표 3. 트리 깊이별 변수 범위
Table 3. Range of variables by tree depth

Depth	5	4
Lux	> 87.5	> 87.5
Temp	> 26.183	> 26.183
Humidity	> 53.338	> 53.338
CO2	> 412.5	-
value	1.35	1.332
squared error	0.004	0.012
sample	1,426	2,453

향후 시간의 흐름에 따른 밀싹의 생육 환경 및 이에 따른 이소오리엔틴 함량 데이터를 수집하면 최적의 생육환경을 찾고 신뢰도 높은 이소오리엔틴 함량을 예측할 수 있는 모형을 만들 수 있을 것으로 예상한다.

References

[1] F. F. Haque, A. Abdelgawad, V. P. Yanambaka, and K. Yelamarthi, "Crop Yield Analysis Using Machine Learning Algorithms", 2020 IEEE 6th World Forum on Internet of Things(WF-IoT), pp. 1-2, Jun. 2020. https://doi.org/10.1109/WF-IoT48130.2020.9221459.

[2] Juan Wang, Jian-li Li, Jing Li, Jin-xin Li, Shu-jie Liu, Lu-qi Huang, and Wen-yuan Gao, "Production of Active Compounds in Medicinal Plants: From Plant Tissue Culture to Biosynthesis", Chinese Herbal Medicines, Vol. 9, No. 2, pp. 115-125, Apr. 2017. https://doi.org/10.1016/S1674-6384(17)60085-6.

[3] S. H. Zendehbad, M. J. Mehran, and Sudhakar Malla, "Flavonoids and phenolic content in wheat grass plant (Triticum aestivum)", Asian Journal of Pharmaceutical and Clinical Research, Vol. 7, No. 4, pp. 184-187, Jan. 2014. https://innovareacademics.in/journals/index.php/ajpcr/article/view/1564.

[4] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques", Computers and Electronics in Agriculture, Vol. 121, pp. 57-65, Dec. 2015. https://doi.org/10.1016/j.compag.2015.11.018.

[5] Valentini Giorgio and Masulli Francesco,

"Ensembles of Learning Machines", Neural Nets WIRN Vietri-2002, Series Lecture Notes in Computer Sciences, Vol. 2486, pp. 3-22, May 2002. https://doi.org/10.1007/3-540-45808-5_1.

[6] Desboulets Loann, "A Review on Variable Selection in Regression Analysis", Econometrics, Vol. 6, No. 45, Nov. 2018. https://doi.org/10.3390/econometrics6040045.

[7] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models", SN Applied Sciences, Vol. 3, No. 272, Feb. 2021. https://doi.org/10.1007/s42452-021-04148-9.

[8] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, Aug. 2016. https://doi.org/10.1145/2939672.2939785.

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting", Advances in Neural Information Processing Systems, Vol. 30, Dec. 2017. https://dl.acm.org/doi/10.5555/3294996.3295074.

[10] Ho Tin Kam, "Random Decision Forests", Proc. of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, pp. 278-282, Aug. 1995. https://doi.org/10.1109/ICDAR.1995.598994.

[11] L. Breiman, "Random Forests", Machine Learning Vol. 45, pp. 5–32, Oct. 2001. https://doi.org/10.1023/A:1010933404324.

저자소개

허 진 경 (JinKyoung Heo)

2004년 2월 : 조선대학교 전산통계학과(공학박사)
2020년 10월 ~ 현재 : 연세대학교 공학연구원(산학협력단) 연구교수
관심분야 : 빅데이터, 인공지능, 인공신경망, 영상처리, 객체탐지