(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0095566
(43) 공개일자 2023년06월29일

(51) 국제특허분류(Int. Cl.)
G06Q 50/26 (2012.01) *G06F 17/18* (2006.01)
G06N 20/00 (2019.01) *G06N 3/02* (2023.01)
G06Q 10/04 (2023.01)
(52) CPC특허분류
G06Q 50/26 (2013.01)
G06F 17/18 (2013.01)
(21) 출원번호 10-2021-0185173
(22) 출원일자 2021년12월22일
심사청구일자 2021년12월22일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
내셔널 유니버시티 오브 싱가포르
싱가포르 119077 로우어 켄트 리지 로드 21
(72) 발명자
옥용식
서울특별시 성북구 고려대로17가길 64, 104동 702호 (안암동1가, 래미안 안암)
왕, 샤오난
싱가포르 117585 비엘케이 이5 #비-03 4 엔지니어링 드라이브 4
(뒷면에 계속)
(74) 대리인
특허법인 남앤남

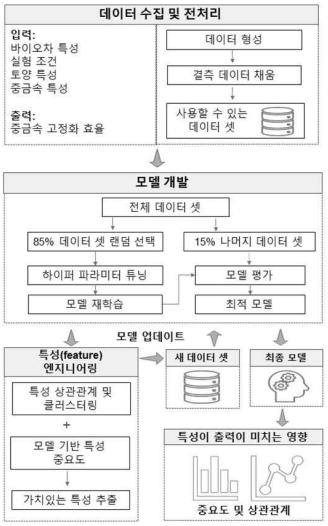
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **토양 중금속 고정 예측 모델 생성 방법 및 장치**

(57) 요 약

토양 중금속 고정 예측 모델 생성 방법 및 장치가 개시된다. 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 방법은, 토양 중금속 고정 예측을 위한 초기 데이터 세트를 수집하는 단계와, 학습 알고리즘이 서로 다른 복수 개의 초기 머신 러닝 모델을 기반으로 초기 데이터 세트에 대한 토양 중금속 고정 예측 학습을 각각 수행하는 단계와, 각각의 학습 결과에 대한 회귀 평가를 기반으로, 복수 개의 초기 머신 러닝 모델 중 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델을 설정하는 단계를 포함할 수 있다.

대 표 도 - 도3



(52) CPC특허분류
G06N 20/00 (2021.08)
G06N 3/02 (2023.01)
G06Q 10/04 (2023.01)

(72) 발명자
쿠무두니, 팔란수리야
서울특별시 성북구 안암로 117 (안암동5가)
원상주
서울특별시 동대문구 무학로49길 58 (용두동)
리, 지에
싱가포르 117585 비엘케이 이5 #비-03 4 엔지니어링 드라이브 4
조유라
서울특별시 성북구 북악산로 742-12 (정릉동, 정릉우정에쉐르)

이 발명을 지원한 국가연구개발사업

과제고유번호	1395068006
과제번호	PJ014758012021
부처명	농촌진흥청
과제관리(전문)기관명	농촌진흥청
연구사업명	미생물활용농업환경문제개선기술개발(R&D)
연구과제명	영농 폐비닐의 작물생산성/환경에 미치는 영향 연구(1주관)
기 여 율	1/2
과제수행기관명	고려대학교(서울)
연구기간	2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711132149
과제번호	2021R1A2C2011734
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	토양 잔류 폐플라스틱의 분해도 및 장기 토양특성 변화평가
기 여 율	1/2
과제수행기관명	고려대학교
연구기간	2021.03.01 ~ 2022.02.28

명 세 서

청구범위

청구항 1

각 단계의 적어도 일부가 프로세서에 의해 수행되는 토양 중금속 고정 예측 모델 생성 방법으로서,

토양 중금속 고정 예측을 위한 초기 데이터 세트를 수집하는 단계;

학습 알고리즘이 서로 다른 복수 개의 초기 머신 러닝 모델을 기반으로 상기 초기 데이터 세트에 대한 토양 중금속 고정 예측 학습을 각각 수행하는 단계; 및

상기 각각의 학습 결과에 대한 회귀 평가를 기반으로, 상기 복수 개의 초기 머신 러닝 모델 중 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델을 설정하는 단계를 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 2

제 1 항에 있어서,

상기 초기 데이터 세트를 수집하는 단계는,

토양 중금속 고정화에 영향을 주는 특성(feature)으로 분류된 입력 데이터와, 상기 입력 데이터에 대응하는 중금속 고정 값을 레이블로 하는 출력 데이터를 수집하는 단계를 포함하며,

상기 토양 중금속 고정화에 영향을 주는 특성은, 열분해 온도와, pH, 표면적(surface area, SA), 조성물(탄소, 수소, 산소, 질소, 재) 및 원자비(H/C, O/C, (O+N)/C)를 포함하는 바이오차 특성과, 바이오차 적용 비율 및 토양 내 바이오차 적용 시간을 포함하는 실험 조건과, 가용 중금속 농도와, 토양 특성 중 적어도 하나 이상을 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 3

제 1 항에 있어서,

상기 초기 데이터 세트에 대한 토양 중금속 고정 예측 학습을 각각 수행하는 단계는,

상기 초기 데이터 세트 중 설정 개수의 학습 데이터를 무작위로 획득하여 상기 서로 다른 복수 개의 초기 머신 러닝 모델을 각각 추론하는 단계를 포함하고,

상기 최적 머신 러닝 모델을 설정하는 단계는,

상기 각각의 학습 결과에 대한 회귀 평가를 기반으로, 실제 값과 예측 값의 정확도가 가장 높은 머신 러닝 모델을 추출하는 단계;

상기 초기 데이터 세트에서 학습 데이터 세트를 제외한 나머지 데이터 세트를 통한 회귀 평가를 기반으로, 상기 추출한 머신 러닝 모델에 대한 검증을 수행하는 단계; 및

상기 검증 결과가 기준 조건을 만족하는 경우, 상기 추출한 머신 러닝 모델을 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델로 설정하는 단계를 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 4

제 3 항에 있어서,

상기 회귀 평가는,

MSE(Mean Squared Error) 및 R^2 중 적어도 하나 이상에 의해 수행되는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 5

제 1 항에 있어서,

상기 초기 데이터 세트에 대하여, 입력 데이터의 특성(feature)의 중요도 및 상기 입력 데이터의 특성간의 상관관계에 기반하여 필터링을 수행하는 단계;

상기 필터링된 데이터 세트를 상기 초기 데이터 세트에 추가하여 업데이트 데이터 세트를 생성하는 단계; 및

상기 업데이트 데이터 세트를 이용하여 상기 설정한 최적 머신 러닝 모델을 업데이트하는 단계를 더 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 6

제 5 항에 있어서,

상기 필터링을 수행하는 단계는,

피어슨 상관관계(Pearson Correlation Coefficient, PCC)를 기반으로 상기 입력 데이터의 특성간의 상관관계를 도출하는 단계; 및

피어슨 순위 상관관계(Pearson Rank Correlation Coefficient)를 기반으로 상기 입력 데이터의 특성에 대한 계층적 군집 분석을 수행하는 단계를 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 7

제 6 항에 있어서,

상기 계층적 군집 분석을 수행하는 단계는,

서로 가장 유사한 두 특성를 추출하기 위해, 기 설정된 거리 임계값을 기반으로 클러스터 수를 결정하여, 특성간 거리가 가장 짧은 특성들을 군집화하는 단계를 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 8

제 7 항에 있어서,

상기 필터링을 수행하는 단계는,

출력 데이터에 대응하는 입력 데이터의 특성의 중요도를 추론하도록 기 학습된 학습 모델에 기반하여, 특성들의 중요도를 판단하는 단계를 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 9

제 8 항에 있어서,

상기 필터링을 수행하는 단계는,

상기 도출한 상관관계가 높은 특성들을 하나의 클러스터로 정렬하는 단계; 및

상기 정렬한 각각의 클러스터 내에서, 상기 특성들의 중요도를 기반으로 대표 특성을 결정하는 단계를 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 10

제 1 항에 있어서,

상기 초기 데이터 세트에서 특성(feature)별 결측 데이터를 예측하여 상기 초기 데이터 세트를 보강하는 단계를 더 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 11

제 10 항에 있어서,

상기 결측 데이터를 예측하는 단계는,

상기 초기 데이터 세트에서 결측 데이터의 특성을 제외한 나머지 특성으로 분류되는 데이터를 상기 설정한 최적 머신 러닝 모델에 입력하는 단계; 및

상기 설정한 최적 머신 러닝 모델의 추론에 따른 출력 값을 결측 데이터로 예측하는 단계를 포함하는,

토양 중금속 고정 예측 모델 생성 방법.

청구항 12

적어도 하나의 프로세서에 의해 실행될 때, 상기 적어도 하나의 프로세서로 하여금 제 1 항 내지 제 11 항 중 어느 하나의 항의 방법을 수행하게 하는 프로그램이 기록된, 컴퓨터에서 판독 가능한 기록 매체.

청구항 13

토양 중금속 고정 예측 모델 생성 장치로서,

메모리; 및

상기 메모리와 연결되고, 상기 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

토양 중금속 고정 예측을 위한 초기 데이터 세트를 수집하는 동작,

학습 알고리즘이 서로 다른 복수 개의 초기 머신 러닝 모델을 기반으로 상기 초기 데이터 세트에 대한 토양 중금속 고정 예측 학습을 각각 수행하는 동작, 및

상기 각각의 학습 결과에 대한 회귀 평가를 기반으로, 상기 복수 개의 초기 머신 러닝 모델 중 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델을 설정하는 동작을 수행하도록 설정되는,

토양 중금속 고정 예측 모델 생성 장치.

청구항 14

제 13 항에 있어서,

상기 초기 데이터 세트에 대한 토양 중금속 고정 예측 학습을 각각 수행하는 동작은,

상기 초기 데이터 세트 중 설정 개수의 학습 데이터를 무작위로 획득하여 상기 서로 다른 복수 개의 초기 머신 러닝 모델을 각각 추론하는 동작을 포함하고,

상기 최적 머신 러닝 모델을 설정하는 동작은,

상기 각각의 학습 결과에 대한 회귀 평가를 기반으로, 실제 값과 예측 값의 정확도가 가장 높은 머신 러닝 모델을 추출하는 동작,

상기 초기 데이터 세트에서 학습 데이터 세트를 제외한 나머지 데이터 세트를 통한 회귀 평가를 기반으로, 상기 추출한 머신 러닝 모델에 대한 검증을 수행하는 동작, 및

상기 검증 결과가 기준 조건을 만족하는 경우, 상기 추출한 머신 러닝 모델을 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델로 설정하는 동작을 포함하는,

토양 중금속 고정 예측 모델 생성 장치.

청구항 15

제 13 항에 있어서,

상기 적어도 하나의 프로세서는,

상기 초기 데이터 세트에 대하여, 입력 데이터의 특성(feature)의 중요도 및 상기 입력 데이터의 특성간의 상관관계에 기반하여 필터링을 수행하는 동작,

상기 필터링된 데이터 세트를 상기 초기 데이터 세트에 추가하여 업데이트 데이터 세트를 생성하는 동작, 및

상기 업데이트 데이터 세트를 이용하여 상기 설정한 최적 머신 러닝 모델을 업데이트하는 동작을 더 수행하도록 설정되는,

토양 중금속 고정 예측 모델 생성 장치.

청구항 16

제 15 항에 있어서,

상기 필터링을 수행하는 동작은,

피어슨 상관관계(Pearson Correlation Coefficient, PCC)를 기반으로 상기 입력 데이터의 특성간의 상관관계를 도출하는 동작, 및

피어슨 순위 상관관계(Pearson Rank Correlation Coefficient)를 기반으로 상기 입력 데이터의 특성에 대한 계층적 군집 분석을 수행하는 동작을 포함하는,

토양 중금속 고정 예측 모델 생성 장치.

청구항 17

제 16 항에 있어서,

상기 계층적 군집 분석을 수행하는 동작은,

서로 가장 유사한 두 특성를 추출하기 위해, 기 설정된 거리 임계값을 기반으로 클러스터 수를 결정하여, 특성간 거리가 가장 짧은 특성들을 군집화하는 동작을 포함하는,

토양 중금속 고정 예측 모델 생성 장치.

청구항 18

제 17 항에 있어서,

상기 필터링을 수행하는 동작은,

출력 데이터에 대응하는 입력 데이터의 특성의 중요도를 추론하도록 기 학습된 학습 모델에 기반하여, 특성들의 중요도를 판단하는 동작을 포함하는,

토양 중금속 고정 예측 모델 생성 장치.

청구항 19

제 18 항에 있어서,

상기 필터링을 수행하는 동작은,

상기 도출한 상관관계가 높은 특성들을 하나의 클러스터로 정렬하는 동작, 및

상기 정렬한 각각의 클러스터 내에서, 상기 특성들의 중요도를 기반으로 대표 특성을 결정하는 동작을

포함하는,

토양 중금속 고정 예측 모델 생성 장치.

청구항 20

제 13 항에 있어서,

상기 적어도 하나의 프로세서는,

상기 초기 데이터 세트에서 특성(feature)별 결측 데이터를 예측하여 상기 초기 데이터 세트를 보강하는 동작을 더 수행하도록 설정되며,

상기 결측 데이터를 예측하는 동작은,

상기 초기 데이터 세트에서 결측 데이터의 특성을 제외한 나머지 특성으로 분류되는 데이터를 상기 설정한 최적 머신 러닝 모델에 입력하는 동작, 및

상기 설정한 최적 머신 러닝 모델의 추론에 따른 출력 값을 결측 데이터로 예측하는 동작을 포함하는,

토양 중금속 고정 예측 모델 생성 장치.

발명의 설명

기 술 분 야

[0001] 본 개시는 바이오차에 의한 토양 중금속 고정을 예측하는 머신러닝 (Machine Learning)을 제안하는 토양 중금속 고정 예측 모델 생성 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 중금속(HM)에 의한 토양오염은 지속 가능한 발전을 위협하는 중대한 세계적 문제이며, 특히 전 세계 인구의 36%가 거주하는 중국과 인도와 같은 개발도상국들의 경우가 그러하다. 채굴 및 제련, 산업 운영, 농업 활동과 같은 인공적 활동은 토양의 중금속오염을 가속화한다.

[0003] 궁극적으로 이러한 중금속은 먹이사슬을 오염시켜 암, 신부전, 심혈관 장애, 신경·인지 장애 등의 질병을 유발하기 때문에, 중금속이 오염된 토양을 교정하기 위해 다양한 현장 및 전 현황 교정 기법을 사용해 왔다.

[0004] 이 중 생물학적 폐기물을 이용한 현장 중금속 고정화는 효율성, 경제성, 적응용이성으로 잘 정착되어 있다. 이 전략은 중금속 생물 가용성이 의도된 토지 사용에 대해 안전한 것으로 간주되는 수준으로 감소하는 위험 기반 교정 전략이다. 적절한 중금속 고정화 계수를 사용하면 환경 설치 공간이 줄어들어 친환경적이고 환경적으로 지속 가능한 문제 해결이 가능할 수 있다.

[0005] 한편 오염된 토양에 바이오차를 적용하는 것은 녹색 개선과 지속 가능한 폐기물 관리 모두를 위한 유망한 전략이다. 바이오차를 사용한 중금속 오염 토양의 교정 조치는 주로 토양, 바이오차 및 중금속 특성에 의존한다. 바이오차 개선 토양에서 중금속 고정화를 위한 최적의 조건을 결정하는 것은 현장마다 다르기 때문에 연구마다 다르다. 따라서 바이오차 개정 토양의 중금속 고정 효율을 예측할 수 있는 일반화된 접근이 필요하다.

[0006] 전술한 배경기술은 발명자가 본 개시의 도출을 위해 보유하고 있었거나, 본 개시의 도출 과정에서 습득한 기술 정보로서, 반드시 본 개시의 출원 전에 일반 공중에게 공개된 공지기술이라 할 수는 없다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 선행기술 1: 한국등록특허 제10-2288644호(2021.08.11.공고)

발명의 내용

해결하려는 과제

[0008] 본 개시의 실시 예의 일 과제는, 복수의 머신러닝(예를 들어, 랜덤 포레스트(RF), 지지 벡터 회귀(SVR), 뉴럴 네트워크(NN))을 이용해 바이오차에 의한 토양 중금속 고정 효율을 예측하여, 토양 중금속 고정화의 극대화를 위한 최적의 실험 조건, 바이오차 특성 및 토양 특성을 결정할 수 있도록 하는데 있다.

[0009] 본 개시의 실시예의 목적은 이상에서 언급한 과제에 한정되지 않으며, 언급되지 않은 본 개시의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 개시의 실시 예에 의해 보다 분명하게 이해될 것이다. 또한, 본 개시의 목적 및 장점들은 특허 청구 범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 알 수 있을 것이다.

과제의 해결 수단

[0010] 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 방법은, 토양 중금속 고정 예측을 위한 초기 데이터 세트를 수집하는 단계와, 학습 알고리즘이 서로 다른 복수 개의 초기 머신 러닝 모델을 기반으로 초기 데이터 세트에 대한 토양 중금속 고정 예측 학습을 각각 수행하는 단계와, 각각의 학습 결과에 대한 회귀 평가를 기반으로, 복수 개의 초기 머신 러닝 모델 중 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델을 설정하는 단계를 포함할 수 있다.

[0011] 이 외에도, 본 개시의 구현하기 위한 다른 방법, 다른 시스템 및 상기 방법을 실행하기 위한 컴퓨터 프로그램이 저장된 컴퓨터로 판독 가능한 기록매체가 더 제공될 수 있다.

[0012] 전술한 것 외의 다른 측면, 특징, 이점이 이하의 도면, 특허청구범위 및 발명의 상세한 설명으로부터 명확해질 것이다.

발명의 효과

[0013] 본 개시의 실시 예에 의하면, 복수의 머신러닝(예를 들어, 랜덤 포레스트(RF), 지지 벡터 회귀(SVR), 뉴럴 네트워크(NN))을 이용하여 바이오차에 의한 토양 중금속 고정 효율을 예측함으로써, 토양 중금속 고정화의 극대화를 위한 최적의 실험 조건, 바이오차 특성 및 토양 특성을 결정할 수 있도록 할 수 있다.

[0014] 또한, 바이오차에 의한 토양 중금속 고정 예측 위한 머신러닝을 이용함으로써, 바이오차 개질 토양에서 중금속 부동화 효율성을 체계적으로 예측할 수 있다.

[0015] 본 개시의 효과는 이상에서 언급된 것들에 한정되지 않으며, 언급되지 아니한 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0016] 도 1은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 시스템을 개략적으로 도시한 도면이다.

도 2는 본 개시의 일 실시 예에 따른 예측 모델 생성 장치를 나타낸 개략적인 블록도이다.

도 3은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 과정을 설명하기 위한 예시도이다.

도 4 및 도 5는 본 개시의 일 실시 예에 따른 결측 데이터 예측을 위한 학습 모델을 설명하기 위한 도면이다.

도 6 및 도 7은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델의 서로 다른 알고리즘의 모델을 비교하기 위한 도면이다.

도 8은 본 개시의 일 실시 예에 따른 데이터 필터링을 위한 입력 특성 분석에 대해 설명하기 위한 도면이다.

도 9는 본 개시의 일 실시 예에 따른 입력 특성의 상관계수를 나타낸 예시도이다.

도 10 및 도 11은 본 개시의 일 실시 예에 따른 업데이트된 토양 중금속 고정 예측 모델을 설명하기 위한 도면이다.

도 12는 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 그래픽 사용자 인터페이스(GUI)의 예시도이다.

도 13은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0017] 본 개시의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 설명되는 실시 예들

을 참조하면 명확해질 것이다.

[0018] 그러나 본 개시는 아래에서 제시되는 실시 예들로 한정되는 것이 아니라, 서로 다른 다양한 형태로 구현될 수 있고, 본 개시의 사상 및 기술 범위에 포함되는 모든 변환, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 아래에 제시되는 실시 예들은 본 개시가 완전하도록 하며, 실시 예들이 속하는 기술분야에서 통상의 지식을 가진 자에게 본 개시의 범주를 완전하게 알려주기 위해 제공되는 것이다. 본 개시를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 개시의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0019] 본 개시에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 개시를 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 개시에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다. 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0020] 이하, 본 개시에 따른 실시 예들을 첨부된 도면을 참조하여 상세히 설명하기로 하며, 첨부 도면을 참조하여 설명함에 있어, 동일하거나 대응하는 구성요소는 동일한 도면번호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다.

[0021] 도 1은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 시스템을 개략적으로 도시한 도면이다.

[0022] 도 1을 참조하면, 토양 중금속 고정 예측 모델 생성 시스템(1)은 예측 모델 생성 장치(100), 사용자 단말(200), 서버(300) 및 네트워크(400)를 포함할 수 있다.

[0023] 일 실시 예에 따른, 토양 중금속 고정 예측 모델 생성 시스템(1)은 바이오차(biochar) 개질 토양에서 중금속 부동화 효율성을 체계적으로 예측하기 위해 머신러닝(Machine Learning, ML) 방식을 채택할 수 있다.

[0024] 이를 통해, 예를 들어, 바이오차(0.3%-25.9% N 범위)와 바이오차 적용률(0.5%-10% 범위)의 질소 함량이 중금속 고정화에 가장 중요한 두 가지 특징 중 하나라는 점을 도출할 수 있다. 또한 인과 분석 결과 중금속 고정 효율성에 있어 중요한 경험 범주는 바이오차 특성 > 실험 조건 > 토양 특성 > 중금속 특성 순이라는 점을 도출할 수 있다.

[0025] 즉 일 실시 예에 따른, 토양 중금속 고정 예측 모델 생성 시스템(1)은 생물학적 특성과 토양 특성이 중금속 고정에 미치는 영향에 대한 새로운 통찰력을 제시하여, 바이오차 개질 토양에서 향상된 중금속 고정화를 위한 최적의 조건을 결정하는 데 도움을 줄 수 있다.

[0026] 오염된 토양에 바이오차를 적용하는 것은 개선된 표면적(SA), 미세공도, 표면 기능군, pH 및 다양한 공급 원료에 대한 양이온 교환 능력을 통해 중금속을 흡착하고 고정할 수 있는 놀라운 능력을 가지고 있기 때문에 중금속 고정에 있어 유망한 방법으로 여겨진다. 예를 들어, 하수 슬러지, 거름, 작물 잔여물은 다양한 생산 조건(예: 느린 열분해, 빠른 열분해 및 기체화)에서, 토양에 비소(As), 카드뮴(Cd), 구리(Cu), 납(Pb), 크롬(Cr), 니켈(Ni), 코발트(Co) 및 아연(Zn)과 같은 중금속(및 준금속)을 고정하는 데 사용되어 왔다.

[0027] 바이오차 개질 토양에서의 중금속 고정 효율은 바이오차 유형(예: 생산 조건, 물리화학적 특성), 토양 특성(예: 토양 pH, 유기 물질 함량, 전기 전도도(EC)) 및 중금속의 특성(예: 밸런스 및 이온 반지름)에 따라 달라질 수 있다. 고정 효율은 다양한 흡착/고정화 메커니즘과 표면 작용기에 의한 양이온 교환, 정전기 상호작용, 침전, 복잡성 등의 요인에 의해 영향을 받을 수 있다.

[0028] 그러나 바이오차를 사용하는 토양에서 중금속 고정화를 강화하기 위한 최적의 조건은 연구 또는 상황에 따라 다를 수 있다. 동시 실험을 통해 공정 모수를 연구하는 것은 어려운 일이기 때문에, 바이오차 처리 토양에서 중금속 고정화 효율을 예측할 수 있는 경험적 접근방식은 토양 중금속 고정화의 극대화를 위해 최적의 실험 조건, 바이오차 특성 및 토양 특성 결정과 관련된 문제를 다룰 수 있어야 한다. 가능한 모든 요인을 포함하는 강력한 모형은 각 요인의 상대적 중요성을 강조할 수 있다. 이를 통해 전체 프로세스에 대한 이해도를 높이고 최적화된 조건에서 오염된 토양에서의 높은 중금속 고정 효율성을 달성할 수 있다.

[0029] 한편, 머신러닝(Machine Learning, ML)은 인공지능의 하위 집합이자 대규모, 복합 및 다차원 데이터를 처리 및

학습하여 예측 모델을 개발할 수 있는 응용 컴퓨팅 기법이다.

[0030] 예를 들어, 랜덤 포레스트(RF), 뉴럴 네트워크(NN) 등의 머신러닝 방식은 토양과 지하수의 오염물질을 감시하고 지도화하는 데 활용될 수 있다. 이러한 머신러닝 모델의 프레임워크는 앙상블 머신러닝 모델링 프로세스를 기반으로 지하수 오염에 대한 위험 평가를 개발할 수 있다는 가능성을 입증할 수 있다. 또한 이 프레임워크는 지하수 오염 위험을 결정하고 건조 지역의 의사결정을 지원하는 데 유용한 정보를 생성하는 데 도움이 될 수 있다. 사용된 머신러닝 모델은 환경 공학 및 생물 교정 조치에 대한 복잡한 시스템의 종속 변수와 독립 변수 간의 비선형 및 복잡한 관계를 예측할 수 있다.

[0031] 다만 바이오차-토양 상호작용의 복잡한 특성 및 체계적인 데이터 세트의 부족으로 인해, 특히 중금속 오염과 관련하여 토양에서 오염물질 고정화를 위한 바이오차 효율의 머신러닝 기반 예측을 수행하는 것에 있어서는 상기 종래의 프레임워크만으로는 예측할 수 없다.

[0032] 이에 일 실시 예에서는, 격차를 해소하기 위해 3가지 머신러닝 모델(예: 랜덤 포레스트(RF), 지지 벡터 회귀(SVR), NN)을 개발하여 바이오차 개질 토양에서 중금속 고정 효율성을 예측할 수 있도록 한다.

[0033] 구체적으로, 일 실시 예에서는, 바이오차 처리 토양에서 중금속 고정화에 미치는 역할과 영향을 평가하기 위해 20개의 입력 변수를 고려할 수 있다. 이 변수들은 네 가지 주요 측면, 즉 i) 바이오차 특성 및 생산 온도, ii) 실험 조건, iii) 바이오차 처리 토양의 물리화학적 특성, 및 iv) 중금속 특성과 관련이 있다.

[0034] 예를 들어, 머신러닝 모델을 생성하기 전에는 문헌에서 수집된 데이터에 대한 철저한 통계 분석을 사용하여 입력 변수(인자)와 출력 변수(중금속 고정)(보조 정보) 사이의 상관 관계를 설정할 수 있다.

[0035] 한편 일 실시 예에서는, 사용자들이 사용자 단말(200)에서 구현되는 어플리케이션 또는 웹사이트에 접속하여, 토양 중금속 고정 예측 모델 생성 과정을 수행할 수 있다. 이러한 사용자 단말(200)은 사용자가 조작하는 데스크 탑 컴퓨터, 스마트폰, 노트북, 태블릿 PC, 스마트 TV, 휴대폰, PDA(personal digital assistant), 랩톱, 미디어 플레이어, 마이크로 서버, GPS(global positioning system) 장치, 전자책 단말기, 디지털방송용 단말기, 네비게이션, 키오스크, MP3 플레이어, 디지털 카메라, 가전기기 및 기타 모바일 또는 비모바일 컴퓨팅 장치일 수 있으나, 이에 제한되지 않는다. 또한, 사용자 단말(200)은 통신 기능 및 데이터 프로세싱 기능을 구비한 시계, 안경, 헤어 밴드 및 반지 등의 웨어러블 단말기 일 수 있다. 사용자 단말(200)은 상술한 내용에 제한되지 아니하며, 웹 브라우징이 가능한 단말기는 제한 없이 차용될 수 있다.

[0036] 한편, 일 실시 예에서, 토양 중금속 고정 예측 모델 생성 시스템(1)은 예측 모델 생성 장치(100) 및/또는 서버(300)에 의해 구현될 수 있는데, 이때 서버(300)는 예측 모델 생성 장치(100)를 운용하기 위한 서버일 수 있다. 즉 서버(300)는 다양한 입력 특성에 따른 토양 중금속 고정화 효율을 예측하는 등의 프로세스를 수행하는 서버일 수 있다.

[0037] 또한 서버(300)는 예측 모델 생성 장치(100)가 구현되도록 하는 데이터를 제공하는 데이터베이스 서버일 수 있다. 그 밖에 서버(300)는 웹 서버 또는 어플리케이션 서버를 포함할 수 있다. 또한 서버(300)는 각종 인공 지능 알고리즘을 적용하는데 필요한 빅데이터 서버 및 AI 서버, 각종 알고리즘의 연산을 수행하는 연산 서버 등을 포함할 수 있으며, 상술하는 서버들을 포함하거나 이러한 서버들과 네트워킹 할 수 있다.

[0038] 네트워크(400)는 토양 중금속 고정 예측 모델 생성 시스템(1)에서 예측 모델 생성 장치(100) 및 서버(300), 그리고 사용자 단말(200)을 연결하는 역할을 수행할 수 있다. 이러한 네트워크(400)는 예컨대 LANs(local area networks), WANs(Wide area networks), MANs(metropolitan area networks), ISDNs(integrated service digital networks) 등의 유선 네트워크나, 무선 LANs, CDMA, 블루투스, 위성 통신 등의 무선 네트워크를 망라할 수 있으나, 본 개시의 범위가 이에 한정되는 것은 아니다. 또한 네트워크(400)는 근거리 통신 및/또는 원거리 통신을 이용하여 정보를 송수신할 수 있다. 여기서 근거리 통신은 블루투스(bluetooth), RFID(radio frequency identification), 적외선 통신(IrDA, infrared data association), UWB(ultra-wideband), ZigBee, Wi-Fi(Wireless fidelity) 기술을 포함할 수 있고, 원거리 통신은 CDMA(code division multiple access), FDMA(frequency division multiple access), TDMA(time division multiple access), OFDMA(orthogonal frequency division multiple access), SC-FDMA(single carrier frequency division multiple access) 기술을 포함할 수 있다.

[0039] 네트워크(400)는 허브, 브리지, 라우터, 스위치 및 게이트웨이와 같은 네트워크 요소들의 연결을 포함할 수 있다. 네트워크(400)는 인터넷과 같은 공용 네트워크 및 안전한 기업 사설 네트워크와 같은 사설 네트워크를 비롯한 하나 이상의 연결된 네트워크들, 예컨대 다중 네트워크 환경을 포함할 수 있다. 네트워크(400)에의 액세스는

하나 이상의 유선 또는 무선 액세스 네트워크들을 통해 제공될 수 있다. 더 나아가 네트워크(400)는 사물 등 분산된 구성 요소들 간에 정보를 주고받아 처리하는 IoT(Internet of Things, 사물인터넷) 망 및/또는 5G 통신을 지원할 수 있다.

[0040] 도 2는 본 개시의 일 실시 예에 따른 예측 모델 생성 장치를 나타낸 개략적인 블록도이다.

[0041] 도 2를 참조하면, 예측 모델 생성 장치(100)는 통신부(110), 사용자 인터페이스(120), 메모리(130) 및 프로세서(140)를 포함할 수 있다.

[0042] 통신부(110)는 네트워크(400)와 연동하여 외부 장치간의 송수신 신호를 패킷 데이터 형태로 제공하는 데 필요한 통신 인터페이스를 제공할 수 있다. 또한 통신부(110)는 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다. 이러한 통신부(110)는 각종 사물 지능 통신(IoT(internet of things), IoE(internet of everything), IoST(internet of small things) 등)을 지원할 수 있으며, M2M(machine to machine) 통신, V2X(vehicle to everything communication) 통신, D2D(device to device) 통신 등을 지원할 수 있다.

[0043] 즉, 프로세서(140)는 통신부(110)를 통해 연결된 외부 장치로부터 각종 데이터 또는 정보를 수신할 수 있으며, 외부 장치로 각종 데이터 또는 정보를 전송할 수도 있다. 그리고, 통신부(110)는 WiFi 모듈, Bluetooth 모듈, 무선 통신 모듈, 및 NFC 모듈 중 적어도 하나를 포함할 수 있다.

[0044] 사용자 인터페이스(120)는 토양 중금속 고정 예측 모델 생성을 위해 데이터 세트를 입력하는 과정, 토양 중금속 고정 예측 모델 및 특성 중요도 산출 모델 등의 파라미터를 설정하는 과정 등 알고리즘 설정 과정 등을 제어하기 위한 사용자 요청 및 명령들이 입력되는 입력 인터페이스를 포함할 수 있다.

[0045] 그리고 사용자 인터페이스(120)는 예측 모델 생성 장치(100)에서 수행된 결과, 즉 예측된 토양 중금속 고정화 효율이 출력되는 출력 인터페이스를 포함할 수 있다. 즉, 사용자 인터페이스(120)는 사용자 요청 및 명령에 따른 결과를 출력할 수 있다. 이러한 사용자 인터페이스(120)의 입력 인터페이스와 출력 인터페이스는 동일한 인터페이스에서 구현될 수 있다.

[0046] 메모리(130)는 예측 모델 생성 장치(100)의 동작 및/또는 서버(300)의 제어(연산)에 필요한 각종 정보들을 저장하고, 제어 소프트웨어를 저장할 수 있는 것으로, 휘발성 또는 비휘발성 기록 매체를 포함할 수 있다.

[0047] 메모리(130)는 하나 이상의 프로세서(140)와 연결되는 것으로, 프로세서(140)에 의해 실행될 때, 프로세서(140)로 하여금 예측 모델 생성 장치(100) 및/또는 서버(300)를 제어하도록 야기하는(cause) 코드들을 저장할 수 있다.

[0048] 여기서, 메모리(130)는 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 일 실시 예에서의 범위가 이에 한정되는 것은 아니다. 이러한 메모리(130)는 내장 메모리 및/또는 외장 메모리를 포함할 수 있으며, DRAM, SRAM, 또는 SDRAM 등과 같은 휘발성 메모리, OTPROM(one time programmable ROM), PROM, EPROM, EEPROM, mask ROM, flash ROM, NAND 플래시 메모리, 또는 NOR 플래시 메모리 등과 같은 비휘발성 메모리, SSD. CF(compact flash) 카드, SD 카드, Micro-SD 카드, Mini-SD 카드, Xd 카드, 또는 메모리 스틱(memory stick) 등과 같은 플래시 드라이브, 또는 HDD와 같은 저장 장치를 포함할 수 있다. 그리고, 메모리(130)에는 일 실시 예에 따른 학습을 수행하기 위한 알고리즘에 관련된 정보가 저장될 수 있다. 그 밖에도 일 실시 예의 목적을 달성하기 위한 범위 내에서 필요한 다양한 정보가 메모리(130)에 저장될 수 있으며, 메모리(130)에 저장된 정보는 서버 또는 외부 장치로부터 수신되거나 사용자에 의해 입력됨에 따라 갱신될 수도 있다.

[0049] 프로세서(140)는 예측 모델 생성 장치(100)의 전반적인 동작을 제어할 수 있다. 구체적으로, 프로세서(140)는 메모리(130)를 포함하는 예측 모델 생성 장치(100)의 구성과 연결되며, 메모리(130)에 저장된 적어도 하나의 명령을 실행하여 예측 모델 생성 장치(100)의 동작을 전반적으로 제어할 수 있다.

[0050] 이러한, 프로세서(140)는 다양한 방식으로 구현될 수 있다. 예를 들어, 프로세서(140)는 주문형 집적 회로(Application Specific Integrated Circuit, ASIC), 임베디드 프로세서, 마이크로 프로세서, 하드웨어 컨트롤 로직, 하드웨어 유한 상태 기계(Hardware Finite State Machine, FSM), 디지털 신호 프로세서(Digital Signal Processor, DSP) 중 적어도 하나로 구현될 수 있다.

[0051] 프로세서(140)는 일종의 중앙처리장치로서 메모리(130)에 탑재된 제어 소프트웨어를 구동하여 예측 모델 생성 장치(100) 전체의 동작을 제어할 수 있다. 프로세서(140)는 데이터를 처리할 수 있는 모든 종류의 장치를 포함

할 수 있다. 여기서, '프로세서(processor)'는, 예를 들어 프로그램 내에 포함된 코드 또는 명령어로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다.

[0053] 도 3은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 과정을 설명하기 위한 예시도이다. 도 3을 참조하여 바이오차 개질 토양의 중금속 고정화 효율을 결정하기 위한 머신러닝 프레임워크에 대해 설명한다.

[0054] 도 3을 참조하면, 일 실시 예에서는 먼저 토양 중금속 고정 예측 모델의 데이터 세트의 구축을 위해, 예를 들어, 네 가지 경험적 범주와 출력 변수를 바탕으로 문헌에서 데이터를 수집할 수 있다. 이때, 일 실시 예에서는, 토양 중금속 고정 예측 모델의 생성을 위해 수집한 데이터 세트의 전처리를 수행할 수 있다.

[0055] 그리고 프로세서(140)는 토양 중금속 고정 예측 모델을 단순화하고 성능을 개선하기 위해 특성(feature) 엔지니어링을 수행하여, 토양 중금속 고정 예측 모델을 업데이트할 수 있다. 이후 일 실시 예에서는, 최종 업데이트된 토양 중금속 고정 예측 모델을 통해, 입력 특성이 출력(토양 중금속 고정화 효율)에 미치는 영향을 판단할 수 있도록 할 수 있다. 이하에서 보다 구체적으로 설명하도록 한다.

[0057] **데이터 세트 컴파일 및 데이터 공백 채우기**

[0058] 일 실시 예에서는, 체계적인 문헌 검토와 데이터 수집을 통해, 20개의 입력 특성에 따른 입력 데이터와, 해당 입력 데이터에 대응하는 출력 변수(중금속 고정화 효율)를 포함하는 데이터 세트를 획득할 수 있다. 이러한 20개의 입력 특성은 도메인 지식을 바탕으로 하여 서로 다른 경험적 범주로 분류될 수 있다.

[0059] 예를 들어, 20개의 입력 특성은 열분해 온도와, pH, 표면적(surface area, SA), 조성물(탄소, 수소, 산소, 질소, 재) 및 원자비(H/C, O/C, (O+N)/C)와 같은 바이오차 특성과, 바이오차 적용 비율 및 토양 내 바이오차 적용 시간과 같은 실험 조건과, 가용 중금속 농도와, 토양 특성을 포함할 수 있다. 그리고 중금속 고정화 효율은 출력 변수로서 획득될 수 있다.

[0060] 한편 데이터 세트를 수집함에 있어, 결측 데이터가 생길 수 있다. 예를 들어, 바이오차 표면적(SA)에서 결측 데이터 포인트가 35개(총 162개 중) 있을 수 있는데, 이때 바이오차 표면적 데이터를 그대로 고려할 경우, 데이터 세트를 더 감소시키게 된다.

[0061] 이에, 일 실시 예에서는, 다른 특성의 입력 데이터들을 통해 결측 데이터를 예측할 수 있다.

[0062] 도 4 및 도 5는 본 개시의 일 실시 예에 따른 결측 데이터 예측을 위한 학습 모델을 설명하기 위한 도면이다.

[0063] 도 4(a), 도 5(a) 및 도 5(b)는 결측 데이터 예측을 위한 학습 모델의 하이퍼 파라미터 조정에 대한 도면이고, 도 4(b), 도 5(c) 및 도 5(d)는 결측 데이터 예측을 위한 학습 모델의 예측 성능을 나타낸 도면이며, 도 4(c)는 표면적 예측 측면에서 결측 데이터 예측을 위한 학습 모델(예를 들어, 랜덤 포레스트 모델)의 입력 데이터 특성별 중요도를 나타낸 도면이다.

[0064] 이때, 결측 데이터 예측을 위한 학습 모델은, 예를 들어, 바이오차 특성(pH, C, H, O, N 함량, H/C, O/C, (O+N)/C, 회분 함량) 및 바이오차 열분해 온도를 포함한 131개의 데이터 포인트를 기반으로 설정될 수 있다.

[0065] 일 실시 예에서는, 데이터 세트의 균일성을 보장하고 결측 데이터 포인트를 채우기 위해, 열분해 온도, 바이오차의 pH, 바이오차의 조성물 및 원자비를 입력으로 사용하여 결측 데이터(표면적)를 채우기 위해, 랜덤 포레스트(이하, RF), 지지 벡터 회귀(이하, SVR), 뉴럴 네트워크(이하, NN)을 포함한 3개의 머신러닝(이하, ML) 알고리즘을 사용할 수 있다.

[0066] 도 4(a)를 참조하여 RF의 하이퍼 파라미터 조정에 대해 확인할 수 있으며, 도 5(a)를 참조하여 SVR의 하이퍼 파라미터 조정에 대해 확인할 수 있고, 도 5(b)를 참조하여 NN의 하이퍼 파라미터 조정에 대해 확인할 수 있다.

[0067] 그리고 도 4(b), 도 5(c) 및 도 5(d)를 참조하여, 세 개의 ML 모델의 표면적에 대한 실제 값(Actual values) 대 예측 값(Predicted values)에 있어, 모든 모델이 높은 검정 결정 계수(R^2 ~ 0.98-0.99)를 사용하여 표면적 예측에 대해 양호한 성능을 나타냈음을 확인할 수 있다.

[0068] 한편, 일 실시 예에서는, RF 모델을 활용하여 결측 값을 예측하고 각 특성 별 중요성을 파악할 수 있다. 도 4(c)는 바이오차 표면적(SA)에 대한 각 특성의 중요성을 보여준다. H/C 원자비는 SA 예측에 있어 가장 중요한

특성임을 확인할 수 있다. 바이오차 표면적에 대해 두 번째로 중요한 특성은 바이오차 pH였으며, 그 다음으로 바이오차 열분해 온도가 높았다. 일반적으로 열분해 온도는 바이오차에서 표면적 및 H/C 원자비와의 강한 상관관계를 보여준다. 예를 들어, 높은 열분해 온도는 표면적을 높이고 바이오차의 H/C 원자비를 낮춘다.

[0069] 즉, 일 실시 예에서는, 예를 들어, 20개의 입력 특성과 출력 변수를 포함한 5개의 경험적 범주에 대한 데이터(총 162개의 데이터 포인트)가 수집되어 ML 학습에 사용될 수 있다. 20개의 매개 변수를 입력 기능으로 식별하고 중금속 고정화 효율은 출력 변수로 정의할 수 있다.

[0070] 이때 총 데이터 세트에서 바이오차 표면적 데이터와 토양 전기전도도에 대한 데이터 포인트가 각각 35개와 10개 누락될 수 있다. 일 실시 예에서는, 이러한 데이터 세트의 격차를 메우고 균일한 데이터 세트를 만들기 위해 pH, 조성물 및 원자비 등 바이오차의 다른 특성을 사용하여 누락된 표면적을 예측하기 위해 ML 모델을 사용할 수 있다. 한편, 토양 전기전도도의 결측 데이터 포인트는 전기전도도를 예측할 토양 입력 변수가 부족하기 때문에 제거될 수 있다.

[0071] 다시 말하면, 일 실시 예에서는, 바이오차 표면적과 고정 효율성 예측을 위해 두 개의 데이터 세트가 컴파일 될 수 있다. 표면적 예측을 위한 데이터 세트에는 바이오차 특성에 대한 131개의 데이터 포인트가 입력 변수로 포함될 수 있고, 고정 효율성 예측을 위한 데이터 세트는 표면적에 대한 결측 데이터를 추가한 후 156개의 데이터 포인트로 구성될 수 있다.

[0073] **ML 모델 생성을 위한 특성(feature) 분석**

[0074] 도 6 및 도 7은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델의 서로 다른 알고리즘의 모델을 비교하기 위한 도면이다.

[0075] 도 6은 바이오차 개질 토양에서의 중금속 고정화 효율을 예측하는 서로 다른 알고리즘의 3 개의 모델의 성능을 나타내는 것으로, 도 6(a)는 랜덤 포레스트(RF), 도 6(b)는 지지 벡터 회귀(SVR) 및 (c) 뉴럴 네트워크(NN)를 나타낸다. 일 실시 예에서는, RF, SVR 및 NN을 실시 예로 하여 설명하고 있으나, 이에 한정되지 않고 다른 알고리즘으로 구현되는 모델이 사용될 수 있다.

[0076] 일 실시 예에서는, 토양 중금속 고정 예측 모델의 학습을 위해 상술한 바와 같이, 바이오차 표면적(SA)의 데이터 격차를 메운 후 총 156개의 데이터 포인트를 획득할 수 있다.

[0077] 일 실시 예에서는, 데이터 간극을 메운 후, 20개의 입력 특성을 기반으로 바이오차 처리 토양에 대한 중금속 고정화 효율성을 예측하기 위해 상술한 바와 같이, 3개의 ML 알고리즘(SVR, NN, RF)을 사용할 수 있다.

[0078] 도 7에 도시된 바와 같이, 각 모델(SVR, NN, RF)에 대한 최적의 하이퍼 파라미터는 5배 교차 검증을 기반으로 예측 오류를 최소화하기 위해 학습 단계에서 조정될 수 있다. 도 7(a)는 지지 벡터 회귀(SVR), 도 7(b)는 뉴럴 네트워크(NN), 도 7(c)는 랜덤 포레스트(RF)의 하이퍼 파라미터에 대해 나타낸다.

[0079] 또한, 도 6은 세 모델의 중금속 고정 효율성에 대한 실제 값과 예측 값을 도시한 것으로, RF 모델의 경우 학습 및 테스트의 R^2 회귀 평가 결과는 각각 0.95와 0.91이었고, RMSE(근 평균 제곱 오차) 회귀 평가 결과는 각각 7.35%와 10.54%임을 볼 수 있다. SVR과 NN 모델의 경우, 테스트의 R^2 회귀 평가 결과는 각각 0.88과 0.80으로 비슷함을 볼 수 있다.

[0080] 이러한 결과는 최적으로 튜닝된 하이퍼 파라미터를 가진 RF 모델이 중금속 고정화 효율성을 예측하는 데 가장 적합한 알고리즘임을 의미할 수 있다.

[0081] 한편, 일 실시 예에서는, 데이터 세트에 대하여, 입력 데이터의 특성(feature)의 중요도 및 입력 데이터의 특성 간의 상관관계에 기반하여 필터링을 수행할 수 있다. 이하 도 8 및 도 9를 참조하여 구체적으로 설명하도록 한다.

[0082] 도 8은 본 개시의 일 실시 예에 따른 데이터 필터링을 위한 입력 특성 분석에 대해 설명하기 위한 도면이고, 도 9는 본 개시의 일 실시 예에 따른 입력 특성의 상관계수를 나타낸 예시도이다.

[0083] 보다 구체적으로, 도 8(a)는 계층적 군집화 결과를 나타낸 도면이며, 도 8(b)는 RF 모델에서 ML 기반 특성 중요도를 나타낸 도면이다.

[0084] 일 실시 예에서는, ML 모델을 단순화하고 성능을 향상시키려면 ML 기반의 기능 중요도 및 기능 상관관계를 사용하여 기능 필터링을 엔지니어링해야 했다. 예를 들어, 피어슨 상관관계(Pearson Correlation Coefficient, PCC)를 사용하여 특성 간의 상관관계를 파악할 수 있으며, 피어슨 순위 상관 관계를 기반으로 계층적 군집 분석을 수행할 수 있다.

[0085] 일 실시 예에서는, 상관관계가 높은 특성을 하나의 클러스터로 정렬하고 유사한 정보가 포함된 특성을 ML 모델 개발을 위해 이 클러스터의 대표로 선택할 수 있다. ML 기반 특성의 중요도에 따라 클러스터의 대표적인 특징이 결정될 수 있다. 클러스터에서 가장 중요하고 대표적인 특성은 특성 중요도와 특성 상관관계를 결합하여 선택될 수 있다.

[0086] 계층적 클러스터링 전략은 세 단계로 요약할 수 있다. 첫째, 원소 간 거리가 가장 낮은 것으로 확인되었으며, 이는 군집화를 위해 서로 가장 유사한 두 원소를 찾는 것을 의미할 수 있다. 둘째, 두 군집화된 요소는 향후 군집화를 위한 새로운 요소로 간주될 수 있다. 마지막으로, 위의 두 단계를 최종 군집을 얻을 때까지 반복할 수 있다.

[0087] 실시 예에 따라, 거리 임계값을 선택하여 클러스터 수를 결정할 수 있다. 특성 중요도는 출력 변수에 상대적인 입력 특성의 역할을 해석할 수 있도록 학습된 RF 모델을 기반으로 예측될 수 있다.

[0088] 일 실시 예에서는, 계층적 군집 분석 알고리즘을 통해 입력 특성을 군집화 할 수 있으며, 이를 도 8(a)에 도시된 바와 같이 여러 색상과 계층적 수준으로 표시할 수 있으며, 예를 들어, Pearson 순위 상관 관계를 기반으로 군집화 할 수 있다. 이때 일 실시 예에서는, 특성 간의 유사성을 확인하기 위해 거리 임계값을 1로 설정할 수 있다. 즉, 동일한 상위 레이어 분기의 아래쪽 분기(특징 하나에 아래쪽 분기가 하나 있음)가 이 거리 임계값 아래에 있는 하나의 군집으로 식별될 수 있다.

[0089] 이에 일 실시 예에서는, 동일한 클러스터에서 유사한 특성을 결정함으로써 모델을 단순화할 수 있는 대표적인 특성을 추출할 수 있다.

[0090] 한편, 도 8에서 T℃는 열분해 온도, pH_BC는 바이오차의 pH, SA는 바이오차의 표면적을 나타내며, C%, H%, O%, N%는 각각 바이오차의 C, H, O, N 함량, H/C, O/C (O+N)/C는 바이오차의 원자비, Ash %는 바이오차의 재 함량을 나타내고, Time은 토양에서의 중금속 함유 시간, Avail. HM은 토양에서 사용 가능한 중금속 함량, pH_soil은 토양 pH, EC_soil은 토양의 전기 전도도, MW는 중금속 분자량, Electronegativity는 중금속의 전기음성, Ion radius는 중금속의 이온 반지름, Valency는 중금속의 원자가를 나타낸다.

[0091] 일 실시 예에서, 예비 데이터 세트(데이터 필터링 전 데이터 세트)는 ML 모델(특히 RF의 경우)에서 만족스러운 예측 정확도를 나타낼 수 있으나, 덜 중요한 기능과 많은 입력 기능의 동시 사용에 의해, 모델의 일반화 용량을 약화시킬 수 있다.

[0092] 따라서 일 실시 예에서는, ML 기반 특성 엔지니어링을 수행할 수 있으며, 이는 성능 향상을 위해 ML 모델을 단순화할 수 있도록 덜 중요한 특성을 필터링하기 위해 수행될 수 있다.

[0093] 일 실시 예에서, 관련성이 높은 입력 데이터 특성은 Pearson 상관 계수(PCC, 도 9 참조)와 계층적 클러스터링(도 8(a) 참조)을 사용하여 식별될 수 있다.

[0094] 클러스터링 및 경험 범주는 도메인 지식을 기반으로 식별될 수 있는데, 여기에는 열분해 온도, 바이오차 특성, 실험 조건, 토양 특성 및 중금속 특성이 포함될 수 있다. 일 실시 예에서는, 동일한 클러스터 내에서 서로 다른 경험적 범주에 속하는 특성은 특성 필터링 과정에서 제거하지 않을 수 있다.

[0095] 도 8(a)를 참조하면, (O+N)/C, O/C 및 O 콘텐츠가 거리 임계값이 1인 한 클러스터 내에 포함되어 있음을 확인할 수 있다. 이러한 특성은 바이오차 특성을 나타내며, ML 모델을 단순화하기 위해 두 가지는 제거할 수 있다. 예를 들어, 도 8(b)의 RF 모델의 특징 중요도를 참조하면, O 콘텐츠는 다른 두 가지((O+N)/C, O/C) 특성보다 훨씬 더 중요한 것으로 파악되므로, (N+O)/C와 O/C는 중복 특성으로 판단하여 제거할 수 있다.

[0096] 또한, 생물학적 특성을 나타내는 H/C와 H 함량도 하나의 클러스터로 식별됨을 확인할 수 있는데, H/C 비율은 H 함량보다 특성의 중요도가 낮기 때문에 제거할 수 있다.

[0097] 마지막 클러스터에는 질량, 전기음성, 이온 반지름 및 밸런스 특성이 포함되어 있는데, 이러한 특성은 모두 중금속 특성이며 상관 관계가 매우 높음을 확인할 수 있다. 이는 이러한 특성이 토양 중금속 고정 예측 모델 생성에 있어 유사한 기여를 제공하므로 중금속 고정 예측에 가장 중요한 특성(전기 음성도)이 모델 생성에 남아 있

음을 나타낼 수 있다.

[0099] **ML 모델 생성 및 평가**

[0100] 일 실시 예에서는, 빠른 수렴을 위한 ML 모델의 학습 과정을 개선하기 위해 입력 특성에 대해 정규화하여 비교적 비슷한 척도와 정규 분포를 구했다. 예를 들어, Scikit-Learn의 StandardScaler가 입력 특성을 표준화 하기 위해 사용될 수 있다.

[0101] 일 실시 예에서는, 정규화 후 데이터 세트를 랜덤하게 둘로 나눌 수 있으며, 이때 ML 모델 학습에 85%가 사용되고, 나머지 15%가 최종 모델 평가에 사용될 수 있다.

[0102] 일 실시 예에서는, 토양 중금속 고정 예측 모델을 학습하기 위해, SVR, RF, NN과 같은 ML 알고리즘을 이용할 수 있다.

[0103] 프로세서(140)는 바이오차 표면적에 대한 최소 평균 제곱 오차를 얻고 5배 교차 검증을 기반으로 고정화 효율성을 예측하기 위해 각 알고리즘에 대한 하이퍼 파라미터를 조정할 수 있다. 그리고 프로세서(140)는 서로 다른 ML 알고리즘에서 학습된 모델을 나머지 15%의 데이터 포인트를 기반으로 평가하여 최적의 ML 모델을 결정할 수 있다.

[0104] 하이퍼 파라미터 조정 프로세스 동안 서로 다른 ML 알고리즘은 다양한 하이퍼 파라미터로 구성될 수 있다. SVR에서 엡실론(ε), 커널 함수 및 페널티(α) 매개변수는 튜닝된 하이퍼 파라미터일 수 있고, RF에서 각 나무의 수와 깊이는 튜닝된 두 가지 중요한 파라미터일 수 있다. 그리고 NN의 경우 모델 융합을 개선하기 위해 각 레이어의 숨겨진 레이어와 뉴런 수, 학습률, 활성화 기능, 솔버(solver) 등을 튜닝할 수 있다.

[0105] SVR, NN 및 RF 알고리즘의 세 가지 최적 ML 모델은 하이퍼 파라미터 튜닝에 따라 획득될 수 있다. 이들 모델은 15% 테스트 데이터 세트를 사용하여 평가될 수 있다. 예측 정확도를 비교하기 위해 결정 계수(R^2)와 RMSE를 활용하여 예측 성능을 정량화할 수 있다. R^2 및 RMSE는 아래 수학식 1 및 수학식 2를 사용하여 계산될 수 있다.

수학식 1

[0106]
$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_p^i - y_t^i)^2}{\sum_{i=1}^{N}(y_t^i - y_m)^2}$$

수학식 2

[0107]
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_t^i - y_p^i)^2}{N}}$$

[0108] 여기서, y_p^i 는 출력의 예측 값이며, y_t^i 는 출력의 실제 값(예를 들어, 실험 연구에 대한 문헌에서 수집된 것)이고, N은 학습 또는 테스트 데이터 세트의 데이터 샘플 수를 나타낸다.

[0110] **최종 특성 탐색을 위한 ML 모델 업데이트**

[0111] 도 10 및 도 11은 본 개시의 일 실시 예에 따른 업데이트된 토양 중금속 고정 예측 모델을 설명하기 위한 도면이다.

[0112] 도 10(a)는 필터링된 데이터 세트를 기반으로 업데이트 된 랜덤 포레스트 모델의 예측 성능을 나타내고, 도 10(b)는 필터링된 데이터 세트를 기반으로 업데이트 된 랜덤 포레스트 모델의 경험적 범주의 상대적 중요성과 특징 중요도 예측 성능을 나타낸다. 그리고 도 11은 필터링된 데이터 세트를 기반으로 업데이트된 랜덤 포레스트 모델의 각 입력 특성에 대한 부분 의존도를 나타낸다.

[0113] 일 실시 예에서는, 성능이 가장 우수한 RF 모델을 선택할 수 있으며, 일반화 기능을 높이고 계산 효율성을 높이기 위해, 상술한 바와 같이 필터링을 수행하여 14개의 입력 특성으로 축소할 수 있다.

[0114] 그리고 일 실시 예에서는, 필터링된 데이터 세트를 이용하여 업데이트를 수행할 수 있으며, RF 모델의 하이퍼 파라미터는 초기 모델과 동일하게 설정할 수 있다.

[0115] 일 실시 예에서는, 데이터 세트의 85%가 새로운 RF 모델을 학습하는데 사용될 수 있으며, 나머지 15%가 예측 성능을 평가하는 데 사용될 수 있다. 도 10(a)를 참조하여, 업데이트된 RF 모델의 예측 정확도를 확인할 수 있으며, 이때 학습과 테스트의 R^2는 각각 0.95와 0.92이고, 학습과 테스트의 RMSE는 각각 7.54%와 9.92%로, 필터링 전 데이터 세트로 학습한 RF 모델에 비해 소폭 증가했음을 확인할 수 있다. RF 모델을 단순화하고 성능을 개선하기 위해 많은 특성이 제거되었기 때문에 이러한 RMSE의 최소 증가는 허용될 수 있다. 이러한 결과는 14개의 중요한 특성만 식별된 특성 필터링 프로세스가 만족스러운 ML 모델 성능을 달성하기에 충분했음을 의미할 수 있다.

[0116] 일 실시 예에서는, 특성 필터링 결과에 따라 선택된 중요하고 가치 있는 특성들은 최적화된 ML 모델을 업데이트하기 위해 추가로 채택될 수 있다. 따라서 하이퍼 파라미터는 그대로 유지되고 입력 특성만 업데이트될 수 있다. ML 모델은 새로운 데이터 세트에 적응하기 위해 다시 학습될 수 있다. 업데이트된 ML 모델은 새로운 데이터 세트의 85%를 무작위로 선택하여 학습한 후 15%의 새로운 데이터로 평가될 수 있다.

[0117] 도 12는 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 그래픽 사용자 인터페이스(GUI)의 예시도이다.

[0118] 일 실시 예에서는, 도 12를 참조하여, 특정 토양 유형에 대한 특정 바이오차의 중금속 고정 성능을 평가하기 위해 제공된 정보를 통해 토양 중금속 고정화 효율을 예측할 수 있도록 할 수 있다. 이는 분석 기구에서 바이오차(초기 조성물, 표면적, 바이오차 준비 온도 및 바이오차 pH), 토양(pH 및 전기 전도도), 중금속(사용 가능한 중금속 농도, 전자 음성도)의 특성이 결정되고 실험에서 개질 조건(바이오차 비율 및 개질 시간)을 기본계획에서 얻으면 달성할 수 있다. 이를 통해 바이오차 개질 토양의 중금속 고정 성능을 조사하는 도메인 연구나 엔지니어링 프로젝트의 시간과 비용을 절약할 수 있다.

[0119] 도 10(b)와 도 11을 참조하여, 업데이트된 RF 모델을 기반으로 출력 중금속 고정 효율성과 관련된 최종 특성의 중요도를 확인할 수 있다. 특히 도 11을 참조하여 바이오차 N 함량과 바이오차 적용률은 고정 효율성을 예측하는 데 가장 중요한 두 가지 특성임을 확인할 수 있다. 도 11을 참조하여, 데이터 범위(0.3%~25.9%) 내에서 N%가 중금속 고정과 양의 상관관계가 있음을 확인할 수 있다.

[0120] 바이오차 표면에 N 함유 기능군(예를 들어, -NH2, N-C=O, C=N)이 존재하면 강한 공유 결합, 수소 결합, 킬레이트화(chelation) 및 정전기 흡인력을 통해 중금속 고정화를 위한 활성 부위가 제공될 수 있다. Cu^{2+}와 같은 중금속은 증가한 -NH2 표면 기능군 때문에 amino-modified 바이오차로 고정될 수 있다. 또한 N 함량이 높은 바이오차는 N 함량이 낮은 바이오차에 비해 흡착 특성이 우수할 수 있다. 예를 들어, N-도핑 바이오차는 순수 바이오차(61.0 $m^2 \cdot g^{-1}$) 보다 높은 SA(418.7 $m^2 \cdot g^{-1}$)로 표면 화학을 변화시킬 수 있다. 이러한 강화된 특성으로 인해 N-도핑된 바이오차는 N 함량이 낮았던 청정 바이오차에 비해 수용성 Cu^{2+}와 Cd^{2+}의 흡착 능력이 더 클 수 있다.

[0121] 또한 도 11을 참조하면, 중금속 고정 효율성은 사용된 데이터 세트를 기준으로 바이오차 적용 비율이 특정 범위(0%-10%) 내에서 증가했음을 확인할 수 있다. 현장 실험과 배양 연구를 통해 바이오차 적용률을 높이면 중금속의 생체 가용성이 감소할 수 있다는 것을 확인할 수 있다. 이는 기능 그룹과 토양 pH가 증가하기 때문에 발생할 수 있으며, 불용성 Cd 침전물은 바이오차 적용률의 증가에 따라 형성될 수 있다.

[0122] 바이오차 개질 토양의 바이오 가용 Pb와 Cd는 바이오차 적용률(0.0%, 1.0%, 2.5%, 5.0%, 10.0%)이 증가함에 따라 감소할 수 있다. 이는 토양 pH(6.17에서 7.17로) 증가와 수용성이 낮은 Pb(OH)2와 (Cd(OH)2) 침전물 형성을

촉진하는 유기물 함량(10.34%에서 11.48%)이 증가하여 Mn 및 Fe 산화물에 결합된 Pb/Cd가 증가했기 때문이다.

[0123] 바이오차 및 토양 EC의 C 함량은 중금속 고정 효율성을 예측하는 데 세 번째와 네 번째로 가장 중요한 특성임을 확인할 수 있으며, 일반적으로 바이오차의 C 함량은 열분해 온도에 따라 증가하는데, 이는 고온에서 생성된 바이오차는 저항성 C가 더 강한 반면 저온에서 생성된 바이오차는 친화성 C이 더 높다는 것을 나타낸다. 특히 저항성 C 이상이 강한 바이오차는 표면 음성도가 커 중금속 고정 성능이 더 우수할 수 있다. 단, 바이오차 C 함량이 50%~60%인 경우 중금속 고정화가 더 높게 나타날 수 있다(도 11 참조).

[0124] 토양 전기 전도도는 바이오차 개질 토양에서 토양 중금속 고정화와 강한 양의 상관관계를 나타낼 수 있다. 토양에 바이오차의 적용은 토양 전기 전도도를 증가시킬 수 있다. 특히 Ca^{2+}, Mg^{2+}, Na^{2+}, K^{2}와 같은 교환 가능한 기본 양이온의 방출로 인해 이는 이온 교환, 복잡성 및 공침을 통한 중금속 고정화를 촉진할 수 있다.

[0125] 비록 중금속 고정에는 바이오차 및 토양의 pH 값이 그다지 중요하게 고려되지 않았지만(도 10(b) 참조), 이러한 특성은 바이오차 적용 후 토양 중금속 고정에 매우 중요할 수 있다. 그러나 도 11을 참조하면, 부분 의존도 그림은 토양 pH와 바이오차 pH가 각각 ~8과 > 7.5일 때 토양 중금속 고정이 더 높음을 확인할 수 있다. 토양 pH가 높을 때 바이오차 표면의 산성 기능군 탈양성은 양전하를 띤 고정화를 흡착하는 선호도를 보여 중금속 고정을 강화할 수 있다. 또한 높은 pH에서 K^{2}, Na^{+}, OH^{-}, PO_{4}^{3-}, Cl^{-} 이온이 방출되어 중금속과 함께 안정적인 화합물 형성이 용이할 수 있다.

[0126] 안정적인 내부 구 복합체의 형성과 탄산염 및 산화물과 같은 중금속의 침전은 바이오차 적용으로 인한 pH 증가로 인해 토양에서 중금속 고정을 촉진하는 추가 메커니즘 중 일부일 수 있다. 특히 바이오차의 표면적과 열분해 온도는 중금속 고정화에 큰 역할을 하지 않았다. 이는 예비 분석과 일치하며, 결과는 표면적의 다른 범위에 높은 고정 효율성이 있고 항상 높은 표면적이 요구되는 것은 아니라는 것을 의미할 수 있다. 오염된 토양에 적용된 바이오차는 복잡한 시스템을 형성하며, 다른 요소들도 바이오차 표면적과 상호작용하여 중금속 고정화에 영향을 미칠 수 있다.

[0127] 한편, 열분해 온도에 따라 바이오차의 표면적이 증가할 수도 있는데, 여기서 표면적이 높은 바이오차는 표면적이 낮은 바이오차 보다 더 많은 중금속을 흡착할 수 있다. 그러나 이러한 현상은 보편적이지 않을 수 있으며 다른 유형의 바이오차, 중금속 종 및 실험 조건에 노출될 경우 변경될 수 있다. 예를 들어, 소나무 톱밥 바이오차(SA = 0.39-0.49 m^{2}/g)와 비교하여 해양 매크로를 이용한 바이오차(SA = 364.47 m^{2}/g)에서 Cu^{2+} 흡착 용량이 더 높을 수 있는데, 이는 특정 표면적이 소나무 톱밥 바이오차 보다 훨씬 낮았음에도 불구하고 나타날 수 있다. 이러한 결과는 바이오차를 사용한 중금속 고정화에서 표면적의 감소된 중요성과 열분해 온도를 강조하는 일 실시 예의 결과를 보완해줄 수 있다.

[0128] 도 11을 참조하면, 바이오차의 회분 함량 증가는 중금속 고정화를 증가시켰음을 확인할 수 있으며, 바이오차의 회분은 용해되면 토양에 양이온을 공급할 수 있는 잔류 미네랄(예: 소금)로 구성될 수 있다. 이러한 양이온은 토양 양이온 교환 용량을 증가시켜 토양의 중금속 고정 효율성을 증가시킬 수 있다.

[0129] 즉, 각 경험적 범주의 상대적 중요도에 기초하여, '바이오차 특성'은 중금속 고정화에 가장 중요한 특성일 수 있으며, 그 다음으로 실험 조건, 토양 특성, 중금속 특성(도 10(b) 참조)의 순으로 중요할 수 있다.

[0131] 도 13은 본 개시의 일 실시 예에 따른 토양 중금속 고정 예측 모델 생성 방법을 설명하기 위한 흐름도이다.

[0132] 도 13을 참조하면, S100단계에서, 프로세서(140)는 토양 중금속 고정 예측을 위한 초기 데이터 세트를 수집한다.

[0133] 이때, 프로세서(140)는 토양 중금속 고정화에 영향을 주는 특성(feature)으로 분류된 입력 데이터와, 상기 입력 데이터에 대응하는 중금속 고정 값을 레이블로 하는 출력 데이터를 수집할 수 있다. 여기서, 토양 중금속 고정화에 영향을 주는 특성은, 열분해 온도와, pH, 표면적(surface area, SA), 조성물(탄소, 수소, 산소, 질소, 재) 및 원자비(H/C, O/C, (O+N)/C)를 포함하는 바이오차 특성과, 바이오차 적용 비율 및 토양 내 바이오차 적용 시간을 포함하는 실험 조건과, 가용 중금속 농도와, 토양 특성 중 적어도 하나 이상을 포함할 수 있다.

[0134] S200단계에서, 프로세서(140)는 학습 알고리즘이 서로 다른 복수 개의 초기 머신 러닝 모델을 기반으로 초기 데이터 세트에 대한 토양 중금속 고정 예측 학습을 각각 수행한다.

[0135] 이때, 프로세서(140)는 초기 데이터 세트 중 설정 개수의 학습 데이터를 무작위로 획득하여 서로 다른 복수 개의 초기 머신 러닝 모델을 각각 추론할 수 있다.

[0136] S300단계에서, 프로세서(140)는 각각의 학습 결과에 대한 회귀 평가를 기반으로, 복수 개의 초기 머신 러닝 모델 중 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델을 설정한다.

[0137] 이때 프로세서(140)는 각각의 학습 결과에 대한 회귀 평가를 기반으로, 실제 값과 예측 값의 정확도가 가장 높은 머신 러닝 모델을 추출할 수 있으며, 초기 데이터 세트에서 학습 데이터 세트를 제외한 나머지 데이터 세트를 통한 회귀 평가를 기반으로, 추출한 머신 러닝 모델에 대한 검증을 수행할 수 있다.

[0138] 그리고 프로세서(140)는 검증 결과가 기준 조건을 만족하는 경우, 추출한 머신 러닝 모델을 토양 중금속 고정 예측을 위한 최적 머신 러닝 모델로 설정할 수 있다.

[0139] 일 실시 예에서, 회귀 평가는 MSE(Mean Squared Error) 및 R^2 중 적어도 하나 이상에 의해 수행될 수 있다.

[0140] 여기서 MSE는 실제 값과 예측 값의 차이를 제곱하여 합한 것을 데이터의 개수만큼 나눈 것을 의미한다. 즉, 평균 제곱 오차로, 잔차 제곱의 합(Residual Sum of Squre, RSS)를 데이터 개수로 나눈 것과 같다. 또한 R^2은 분산 기반으로 예측 성능을 평가하는 것으로, 예측 값의 분산을 실제 값의 분산으로 나누어 1에 가까울 수록 예측 정확도가 높다고 판단할 수 있다.

[0141] 한편, 일 실시 예에서, 프로세서(140)는 초기 데이터 세트에서 특성(feature)별 결측 데이터를 예측하여 초기 데이터 세트를 보강할 수 있다. 이때 프로세서(140)는 초기 데이터 세트에서 결측 데이터의 특성을 제외한 나머지 특성으로 분류되는 데이터를 설정한 최적 머신 러닝 모델에 입력하고, 설정한 최적 머신 러닝 모델의 추론에 따른 출력 값을 결측 데이터로 예측할 수 있다.

[0142] S400단계에서, 프로세서(140)는 초기 데이터 세트에 대하여, 입력 데이터의 특성(feature)의 중요도 및 입력 데이터의 특성간의 상관관계에 기반하여 필터링을 수행하고, S500단계에서, 필터링된 데이터 세트를 초기 데이터 세트에 추가하여 업데이트 데이터 세트를 생성한다.

[0143] 이때, 프로세서(140)는 피어슨 상관관계(Pearson Correlation Coefficient, PCC)를 기반으로 입력 데이터의 특성간의 상관관계를 도출할 수 있다. 또한 프로세서(140)는 피어슨 순위 상관관계(Pearson Rank Correlation Coefficient)를 기반으로 입력 데이터의 특성에 대한 계층적 군집 분석을 수행할 수 있는데, 서로 가장 유사한 두 특성를 추출하기 위해, 기 설정된 거리 임계값을 기반으로 클러스터 수를 결정하여, 특성 간 거리가 가장 짧은 특성들을 군집화할 수 있다.

[0144] 그리고 프로세서(140)는 출력 데이터에 대응하는 입력 데이터의 특성의 중요도를 추론하도록 기 학습된 학습 모델에 기반하여, 특성들의 중요도를 판단할 수 있다. 이때 프로세서(140)는 도출한 상관관계가 높은 특성들을 하나의 클러스터로 정렬하고, 정렬한 각각의 클러스터 내에서, 특성들의 중요도를 기반으로 대표 특성을 결정할 수 있다.

[0145] 그리고 S600단계에서, 프로세서(140)는 업데이트 데이터 세트를 이용하여 설정한 최적의 머신 러닝 모델을 업데이트한다.

[0146] 즉 일 실시 예에서는, 예측 성능이 우수한 최종 ML 모델을 적용하여 각 특성이 대상에 미치는 영향과 특징의 중요성을 도출할 수 있다. 특성 중요도는 최종 업데이트된 RF 모델을 통해 직접 결정될 수 있다. 부분 의존도를 사용하여 업데이트된 RF 모델을 통합하여 각 특성의 상관관계를 출력 변수에 체계적으로 표현할 수 있다. 예측된 출력에 대한 각 특성의 한계 효과는 ML 모델뿐만 아니라 입력 특성과 출력 변수 사이의 관련성(예: 선형, 단조, 심지어 더 복잡한 관계)을 사용하여 결정될 수 있다.

[0148] 이상 설명된 본 개시에 따른 실시 예는 컴퓨터 상에서 다양한 구성요소를 통하여 실행될 수 있는 컴퓨터 프로그램의 형태로 구현될 수 있으며, 이와 같은 컴퓨터 프로그램은 컴퓨터로 판독 가능한 매체에 기록될 수 있다. 이때, 매체는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매

체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등과 같은, 프로그램 명령어를 저장하고 실행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다.

[0149] 한편, 상기 컴퓨터 프로그램은 본 개시를 위하여 특별히 설계되고 구성된 것이거나 컴퓨터 소프트웨어 분야의 통상의 기술자에게 공지되어 사용 가능한 것일 수 있다. 컴퓨터 프로그램의 예에는, 컴파일러에 의하여 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용하여 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함될 수 있다.

[0150] 본 개시의 명세서(특히 특허청구범위에서)에서 "상기"의 용어 및 이와 유사한 지시 용어의 사용은 단수 및 복수 모두에 해당하는 것일 수 있다. 또한, 본 개시에서 범위(range)를 기재한 경우 상기 범위에 속하는 개별적인 값을 적용한 발명을 포함하는 것으로서(이에 반하는 기재가 없다면), 발명의 상세한 설명에 상기 범위를 구성하는 각 개별적인 값을 기재한 것과 같다.

[0151] 본 개시에 따른 방법을 구성하는 단계들에 대하여 명백하게 순서를 기재하거나 반하는 기재가 없다면, 상기 단계들은 적당한 순서로 행해질 수 있다. 반드시 상기 단계들의 기재 순서에 따라 본 개시가 한정되는 것은 아니다. 본 개시에서 모든 예들 또는 예시적인 용어(예들 들어, 등등)의 사용은 단순히 본 개시를 상세히 설명하기 위한 것으로서 특허청구범위에 의해 한정되지 않는 이상 상기 예들 또는 예시적인 용어로 인해 본 개시의 범위가 한정되는 것은 아니다. 또한, 통상의 기술자는 다양한 수정, 조합 및 변경이 부가된 특허청구범위 또는 그 균등물의 범주 내에서 설계 조건 및 팩터에 따라 구성될 수 있음을 알 수 있다.

[0152] 따라서, 본 개시의 사상은 상기 설명된 실시 예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등한 또는 이로부터 등가적으로 변경된 모든 범위는 본 개시의 사상의 범주에 속한다고 할 것이다.

부호의 설명

[0153] 1 : 토양 중금속 고정 예측 모델 생성 시스템

100 : 토양 중금속 고정 예측 모델 생성 장치

110 : 통신부

120 : 사용자 인터페이스

130 : 메모리

140 : 프로세서

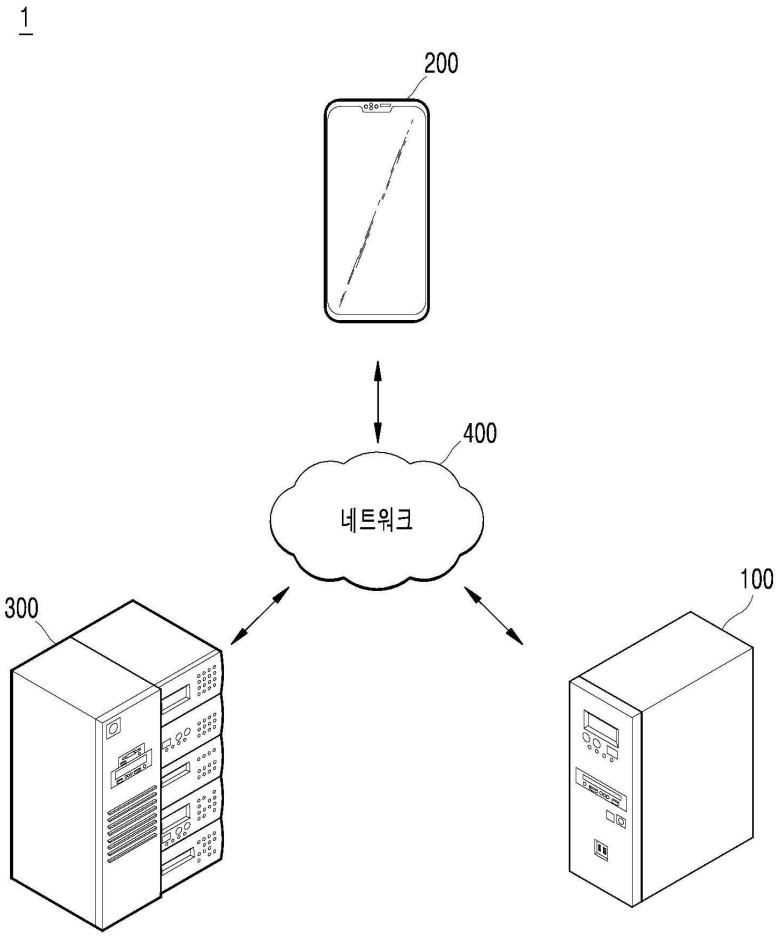
200 : 사용자 단말

300 : 서버

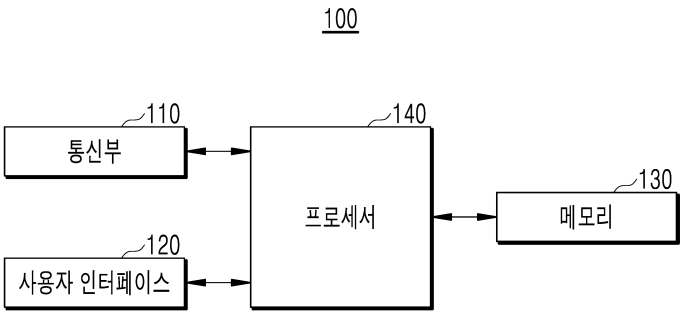
400 : 네트워크

도면

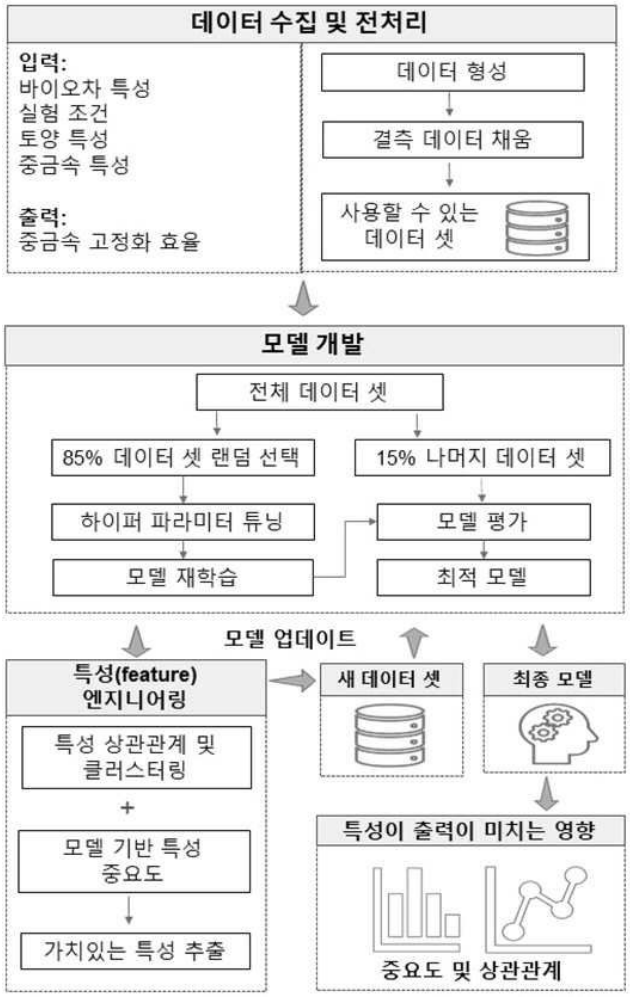
도면1


1
200
400
네트워크
300
100

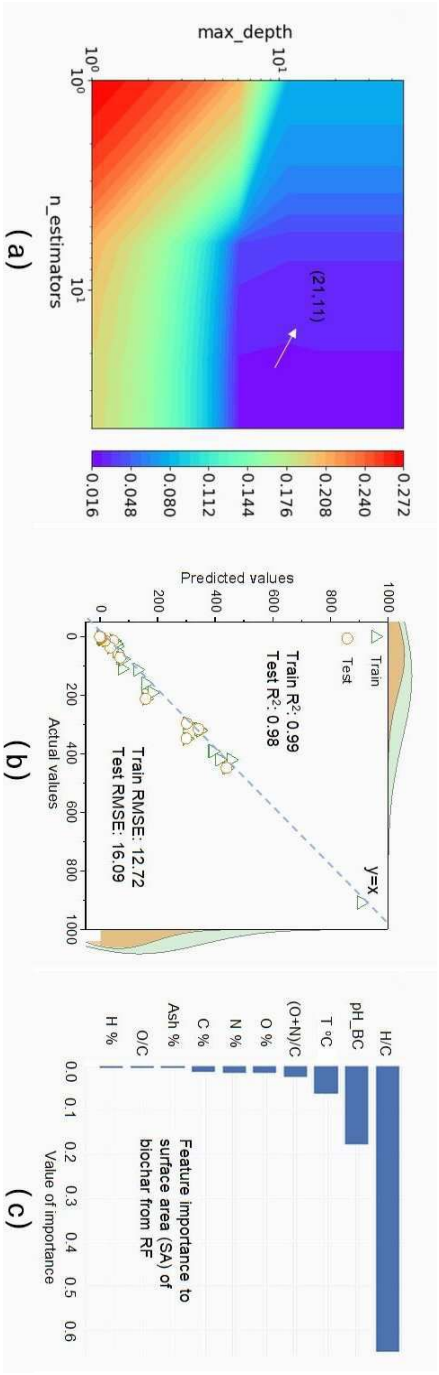
도면2


100
110
통신부
140
프로세서
130
메모리
120
사용자 인터페이스

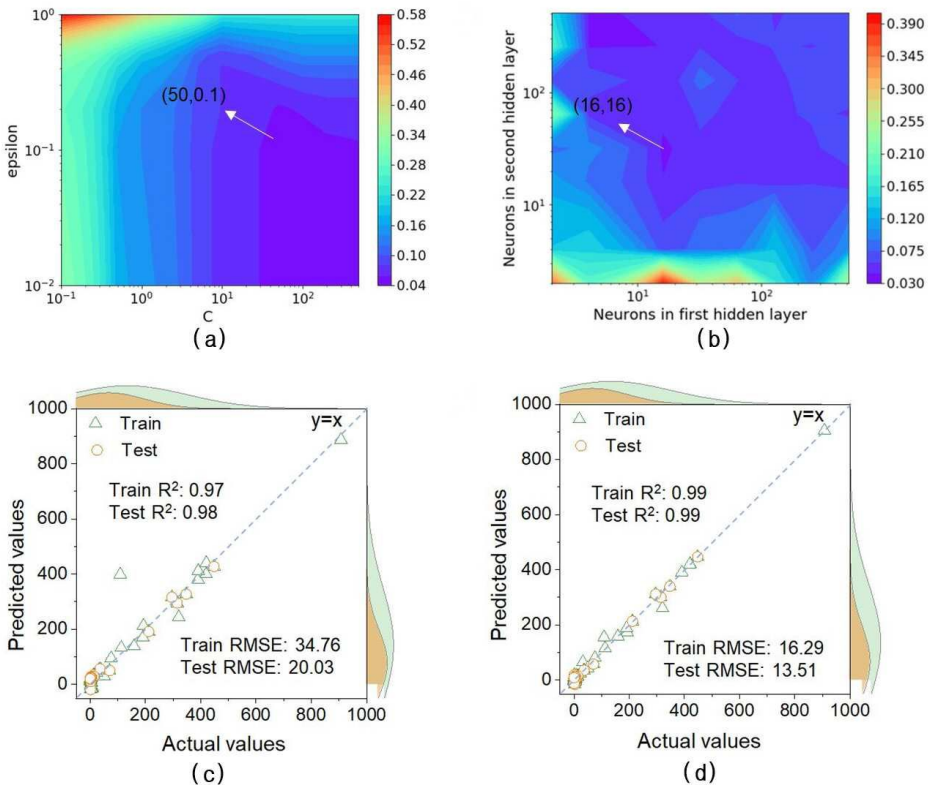
도면3


데이터 수집 및 전처리
입력:
바이오차 특성
실험 조건
토양 특성
중금속 특성
출력:
중금속 고정화 효율
데이터 형성
결측 데이터 채움
사용할 수 있는 데이터 셋
모델 개발
전체 데이터 셋
85% 데이터 셋 랜덤 선택
15% 나머지 데이터 셋
하이퍼 파라미터 튜닝
모델 평가
모델 재학습
최적 모델
모델 업데이트
특성(feature) 엔지니어링
특성 상관관계 및 클러스터링
+
모델 기반 특성 중요도
가치있는 특성 추출
새 데이터 셋
최종 모델
특성이 출력이 미치는 영향
중요도 및 상관관계

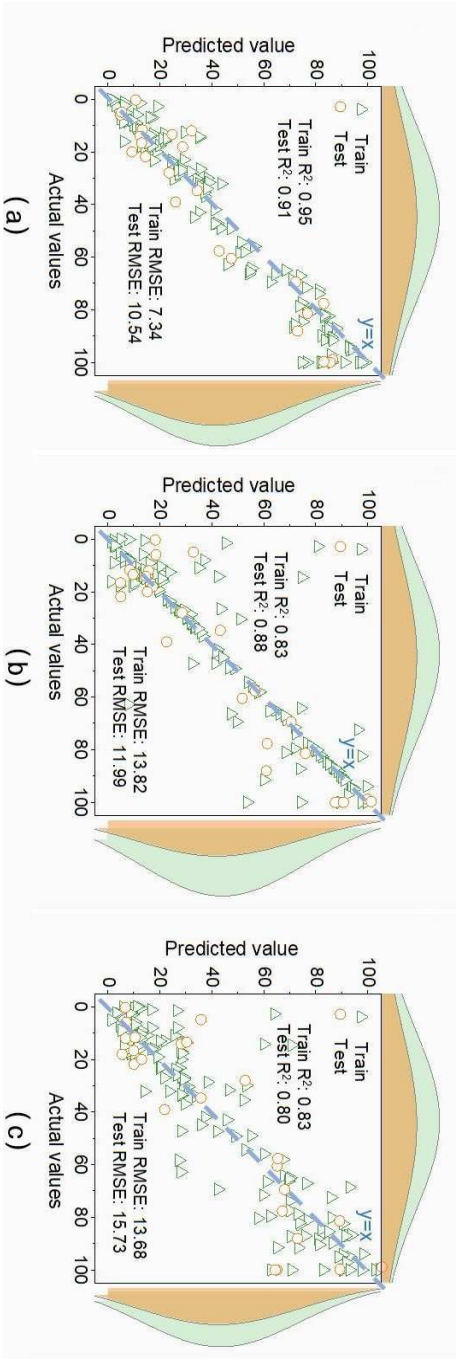
도면4


(a)
max_depth
n_estimators
(21,11)
0.016
0.048
0.080
0.112
0.144
0.176
0.208
0.240
0.272
(b)
Predicted values
Actual values
Train
Test
y=x
Train R²: 0.99
Test R²: 0.98
Train RMSE: 12.72
Test RMSE: 16.09
(c)
H/C
pH_BC
T °C
(O+N)/C
O %
N %
C %
Ash %
O/C
H %
Feature importance to surface area (SA) of biochar from RF
Value of importance

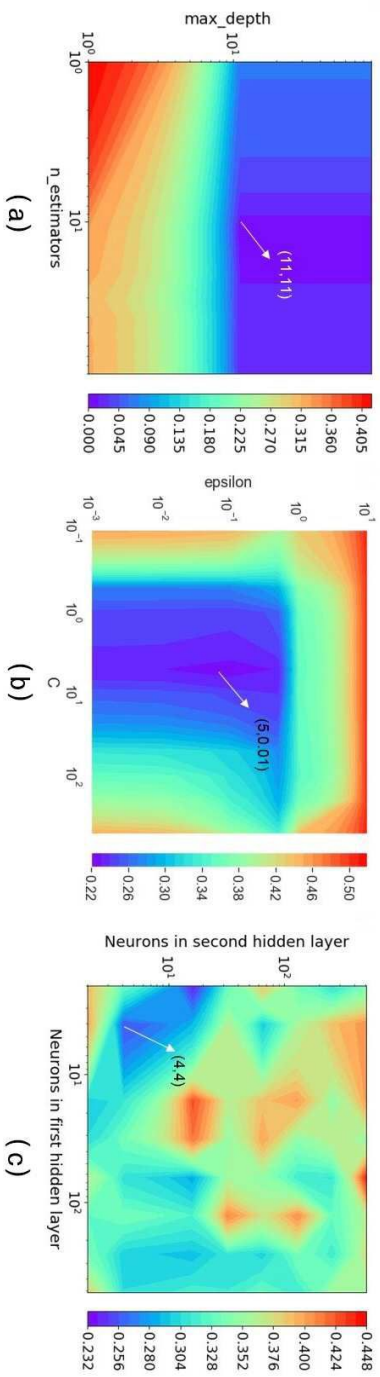
도면5


(50,0.1)
epsilon
C
(a)
(16,16)
Neurons in second hidden layer
Neurons in first hidden layer
(b)
Train
Test
y=x
Train R²: 0.97
Test R²: 0.98
Train RMSE: 34.76
Test RMSE: 20.03
Predicted values
Actual values
(c)
Train
Test
y=x
Train R²: 0.99
Test R²: 0.99
Train RMSE: 16.29
Test RMSE: 13.51
Predicted values
Actual values
(d)

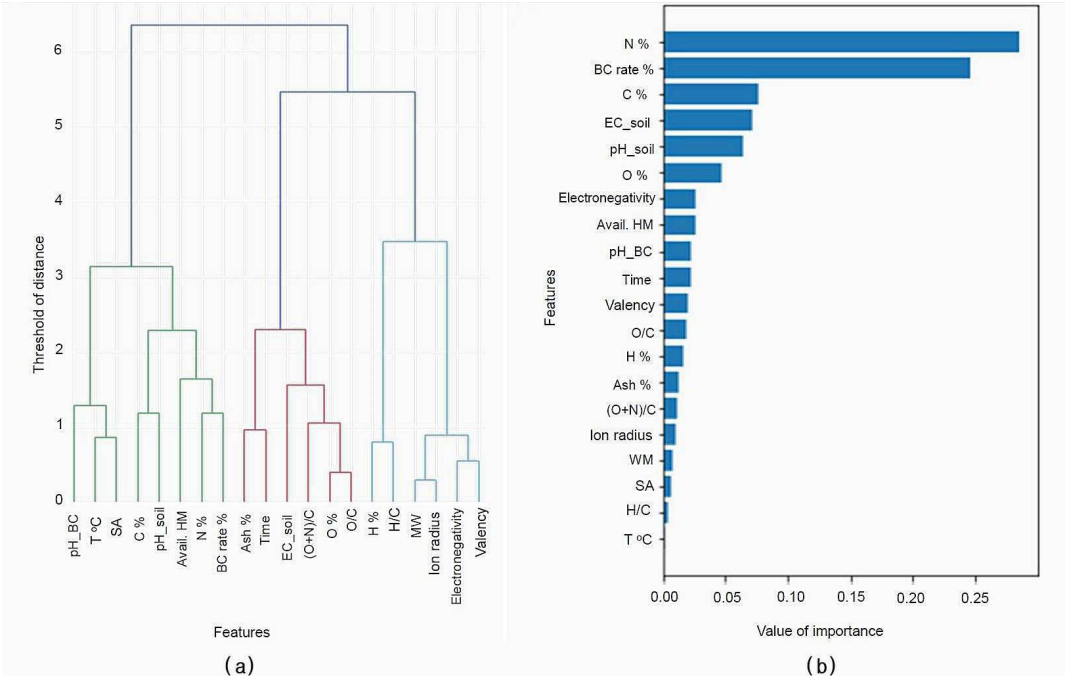
도면6


(a)
Predicted value
Actual values
Train
Test
Train R²: 0.95
Test R²: 0.91
Train RMSE: 7.34
Test RMSE: 10.54
y=x
(b)
Predicted value
Actual values
Train
Test
Train R²: 0.83
Test R²: 0.88
Train RMSE: 13.82
Test RMSE: 11.99
y=x
(c)
Predicted value
Actual values
Train
Test
Train R²: 0.83
Test R²: 0.80
Train RMSE: 13.68
Test RMSE: 15.73
y=x

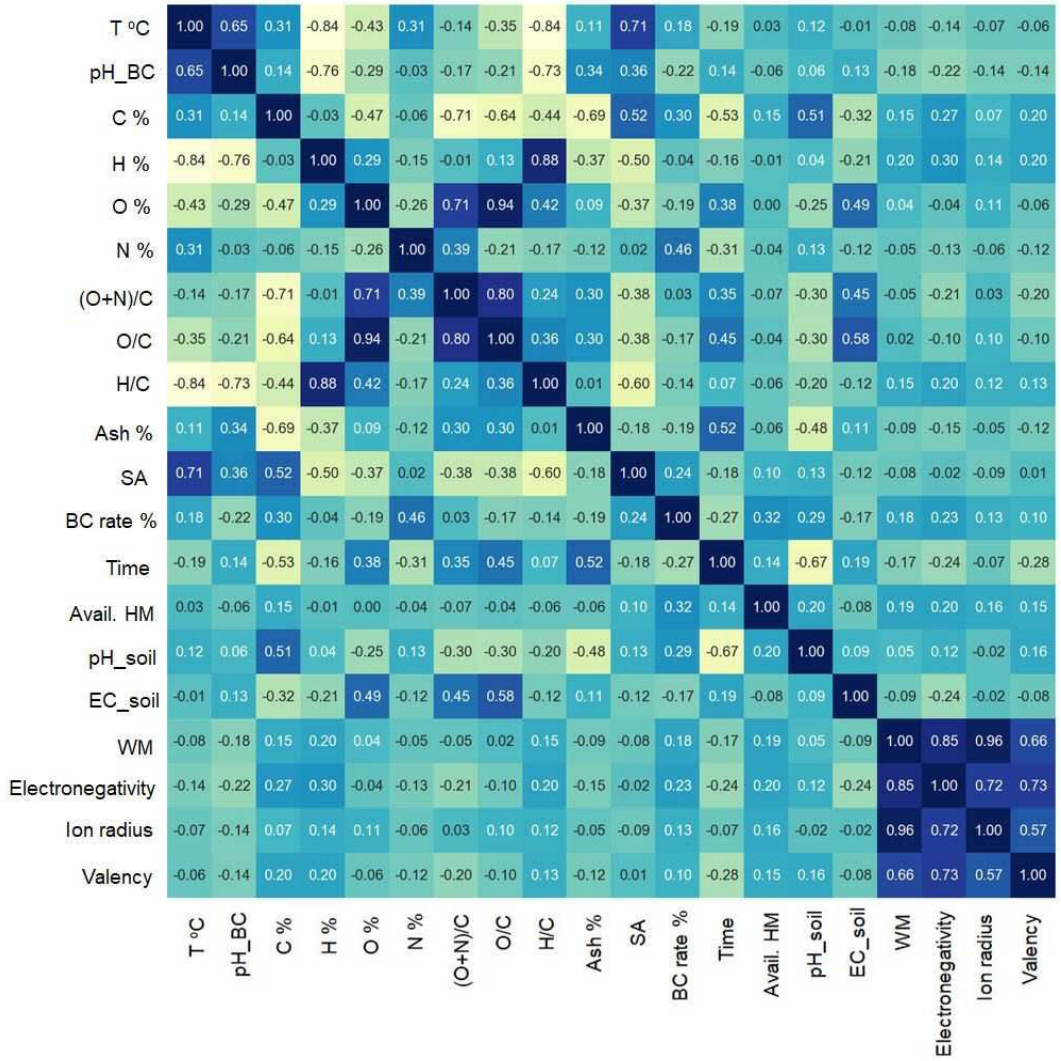
도면7


(a)
max_depth
n_estimators
(11,11)
0.000
0.045
0.090
0.135
0.180
0.225
0.270
0.315
0.360
0.405
(b)
epsilon
C
(5,0.01)
0.22
0.26
0.30
0.34
0.38
0.42
0.46
0.50
(c)
Neurons in second hidden layer
Neurons in first hidden layer
(4,4)
0.232
0.256
0.280
0.304
0.328
0.352
0.376
0.400
0.424
0.448

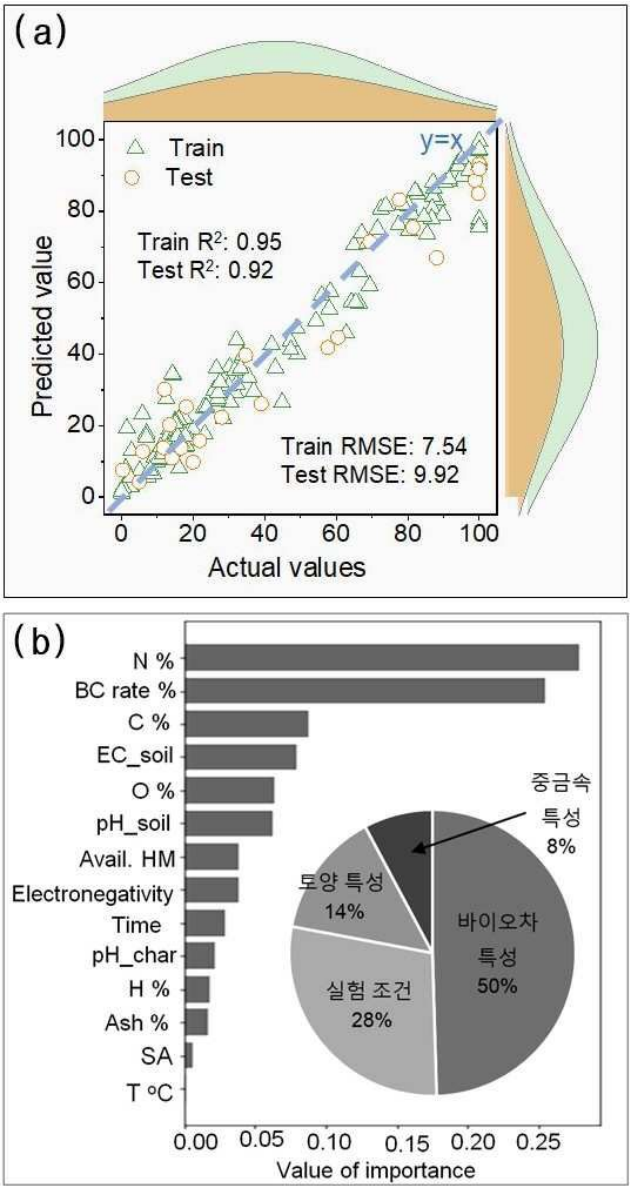
도면8


Threshold of distance
0
1
2
3
4
5
6
pH_BC
T °C
SA
C %
pH_soil
Avail. HM
N %
BC rate %
Ash %
Time
EC_soil
(O+N)/C
O %
O/C
H %
H/C
MW
Ion radius
Electronegativity
Valency
Features
(a)
Features
N %
BC rate %
C %
EC_soil
pH_soil
O %
Electronegativity
Avail. HM
pH_BC
Time
Valency
O/C
H %
Ash %
(O+N)/C
Ion radius
WM
SA
H/C
T °C
0.00
0.05
0.10
0.15
0.20
0.25
Value of importance
(b)

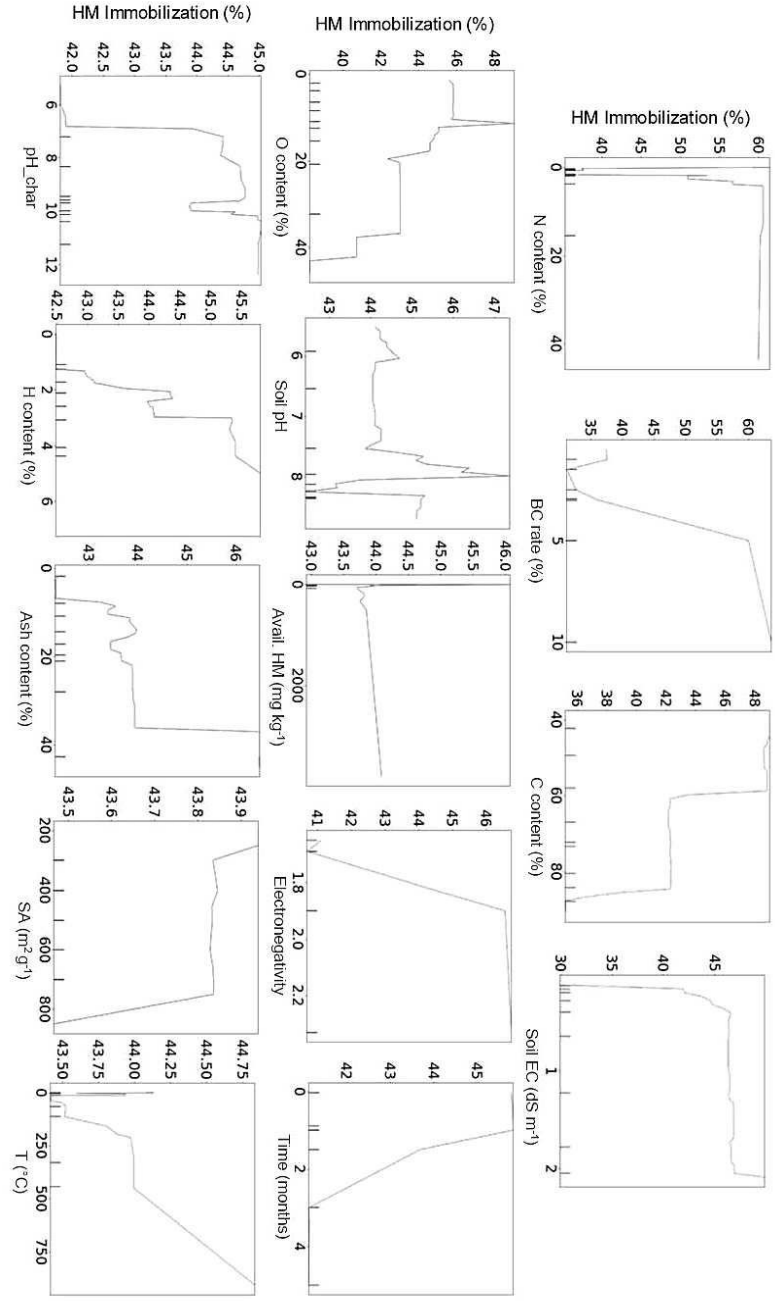
도면9


T °C 1.00 0.65 0.31 -0.84 -0.43 0.31 -0.14 -0.35 -0.84 0.11 0.71 0.18 -0.19 0.03 0.12 -0.01 -0.08 -0.14 -0.07 -0.06
pH_BC 0.65 1.00 0.14 -0.76 -0.29 -0.03 -0.17 -0.21 -0.73 0.34 0.36 -0.22 0.14 -0.06 0.06 0.13 -0.18 -0.22 -0.14 -0.14
C % 0.31 0.14 1.00 -0.03 -0.47 -0.06 -0.71 -0.64 -0.44 -0.69 0.52 0.30 -0.53 0.15 0.51 -0.32 0.15 0.27 0.07 0.20
H % -0.84 -0.76 -0.03 1.00 0.29 -0.15 -0.01 0.13 0.88 -0.37 -0.50 -0.04 -0.16 -0.01 0.04 -0.21 0.20 0.30 0.14 0.20
O % -0.43 -0.29 -0.47 0.29 1.00 -0.26 0.71 0.94 0.42 0.09 -0.37 -0.19 0.38 0.00 -0.25 0.49 0.04 -0.04 0.11 -0.06
N % 0.31 -0.03 -0.06 -0.15 -0.26 1.00 0.39 -0.21 -0.17 -0.12 0.02 0.46 -0.31 -0.04 0.13 -0.12 -0.05 -0.13 -0.06 -0.12
(O+N)/C -0.14 -0.17 -0.71 -0.01 0.71 0.39 1.00 0.80 0.24 0.30 -0.38 0.03 0.35 -0.07 -0.30 0.45 -0.05 -0.21 0.03 -0.20
O/C -0.35 -0.21 -0.64 0.13 0.94 -0.21 0.80 1.00 0.36 0.30 -0.38 -0.17 0.45 -0.04 -0.30 0.58 0.02 -0.10 0.10 -0.10
H/C -0.84 -0.73 -0.44 0.88 0.42 -0.17 0.24 0.36 1.00 0.01 -0.60 -0.14 0.07 -0.06 -0.20 -0.12 0.15 0.20 0.12 0.13
Ash % 0.11 0.34 -0.69 -0.37 0.09 -0.12 0.30 0.30 0.01 1.00 -0.18 -0.19 0.52 -0.06 -0.48 0.11 -0.09 -0.15 -0.05 -0.12
SA 0.71 0.36 0.52 -0.50 -0.37 0.02 -0.38 -0.38 -0.60 -0.18 1.00 0.24 -0.18 0.10 0.13 -0.12 -0.08 -0.02 -0.09 0.01
BC rate % 0.18 -0.22 0.30 -0.04 -0.19 0.46 0.03 -0.17 -0.14 -0.19 0.24 1.00 -0.27 0.32 0.29 -0.17 0.18 0.23 0.13 0.10
Time -0.19 0.14 -0.53 -0.16 0.38 -0.31 0.35 0.45 0.07 0.52 -0.18 -0.27 1.00 0.14 -0.67 0.19 -0.17 -0.24 -0.07 -0.28
Avail. HM 0.03 -0.06 0.15 -0.01 0.00 -0.04 -0.07 -0.04 -0.06 -0.06 0.10 0.32 0.14 1.00 0.20 -0.08 0.19 0.20 0.16 0.15
pH_soil 0.12 0.06 0.51 0.04 -0.25 0.13 -0.30 -0.30 -0.20 -0.48 0.13 0.29 -0.67 0.20 1.00 0.09 0.05 0.12 -0.02 0.16
EC_soil -0.01 0.13 -0.32 -0.21 0.49 -0.12 0.45 0.58 -0.12 0.11 -0.12 -0.17 0.19 -0.08 0.09 1.00 -0.09 -0.24 -0.02 -0.08
WM -0.08 -0.18 0.15 0.20 0.04 -0.05 -0.05 0.02 0.15 -0.09 -0.08 0.18 -0.17 0.19 0.05 -0.09 1.00 0.85 0.96 0.66
Electronegativity -0.14 -0.22 0.27 0.30 -0.04 -0.13 -0.21 -0.10 0.20 -0.15 -0.02 0.23 -0.24 0.20 0.12 -0.24 0.85 1.00 0.72 0.73
Ion radius -0.07 -0.14 0.07 0.14 0.11 -0.06 0.03 0.10 0.12 -0.05 -0.09 0.13 -0.07 0.16 -0.02 -0.02 0.96 0.72 1.00 0.57
Valency -0.06 -0.14 0.20 0.20 -0.06 -0.12 -0.20 -0.10 0.13 -0.12 0.01 0.10 -0.28 0.15 0.16 -0.08 0.66 0.73 0.57 1.00
T °C
pH_BC
C %
H %
O %
N %
(O+N)/C
O/C
H/C
Ash %
SA
BC rate %
Time
Avail. HM
pH_soil
EC_soil
WM
Electronegativity
Ion radius
Valency

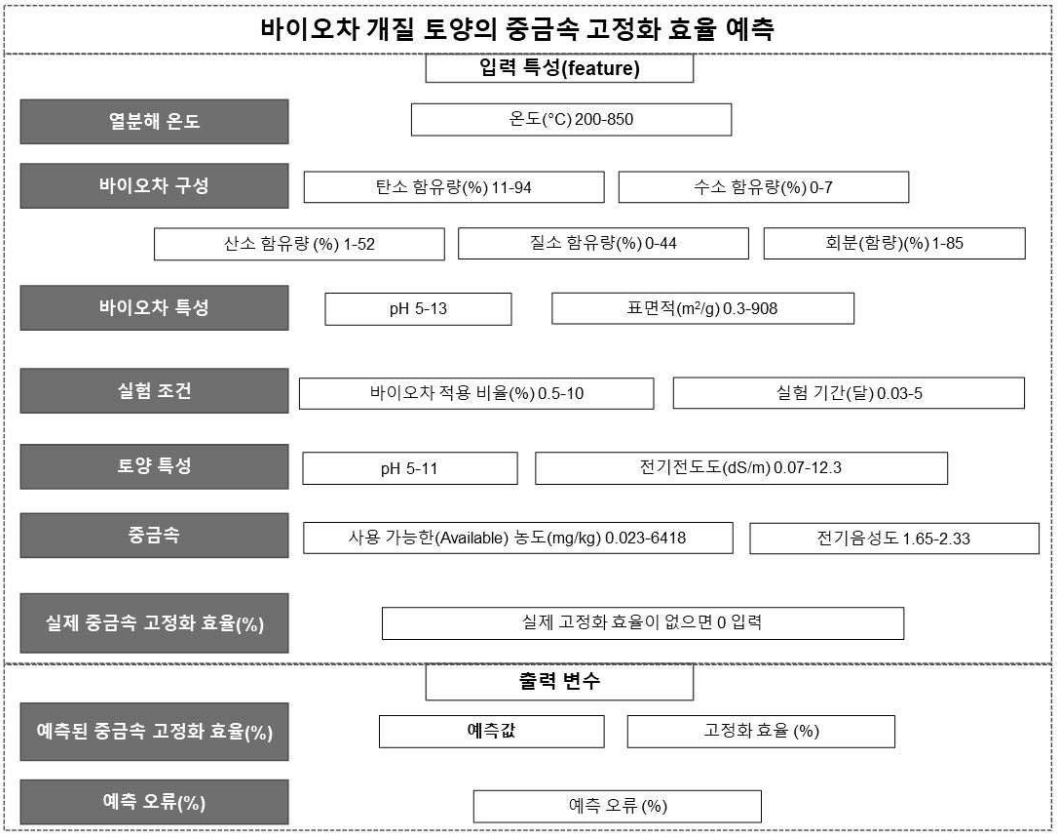
도면10


(a)
Train
Test
y=x
Train R^2: 0.95
Test R^2: 0.92
Train RMSE: 7.54
Test RMSE: 9.92
Predicted value
Actual values
0
20
40
60
80
100
(b)
N %
BC rate %
C %
EC_soil
O %
pH_soil
Avail. HM
Electronegativity
Time
pH_char
H %
Ash %
SA
T °C
0.00
0.05
0.10
0.15
0.20
0.25
Value of importance
중금속 특성 8%
토양 특성 14%
바이오차 특성 50%
실험 조건 28%

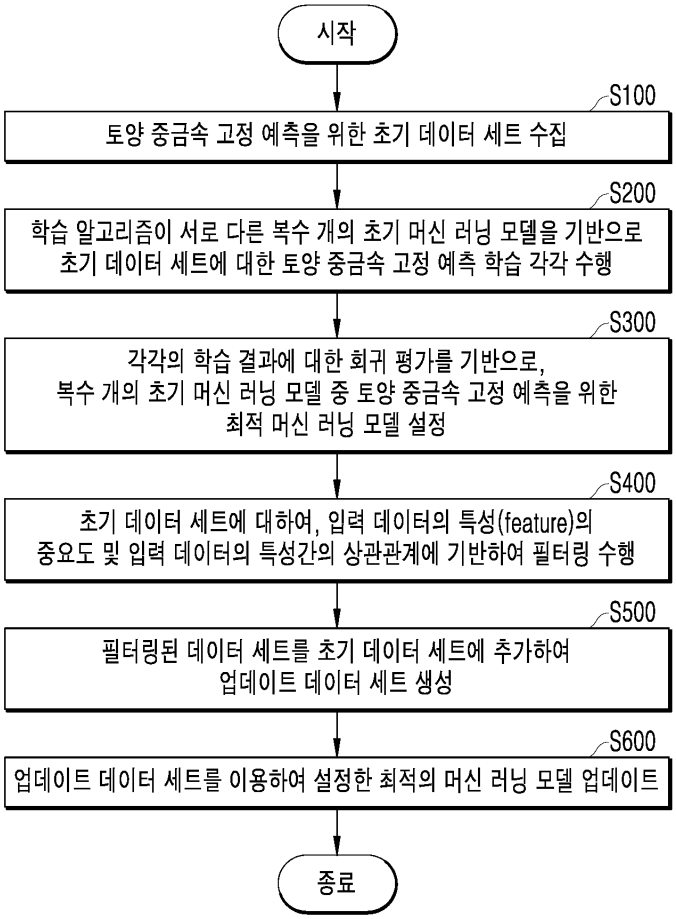
도면11


HM Immobilization (%)
N content (%)
BC rate (%)
C content (%)
Soil EC (dS m-1)
O content (%)
Soil pH
Avail. HM (mg kg-1)
Electronegativity
Time (months)
pH_char
H content (%)
Ash content (%)
SA (m2 g-1)
T (°C)

도면12


바이오차 개질 토양의 중금속 고정화 효율 예측
입력 특성(feature)
열분해 온도
온도(°C) 200-850
바이오차 구성
탄소 함유량(%) 11-94
수소 함유량(%) 0-7
산소 함유량 (%) 1-52
질소 함유량(%) 0-44
회분(함량)(%) 1-85
바이오차 특성
pH 5-13
표면적(m²/g) 0.3-908
실험 조건
바이오차 적용 비율(%) 0.5-10
실험 기간(달) 0.03-5
토양 특성
pH 5-11
전기전도도(dS/m) 0.07-12.3
중금속
사용 가능한(Available) 농도(mg/kg) 0.023-6418
전기음성도 1.65-2.33
실제 중금속 고정화 효율(%)
실제 고정화 효율이 없으면 0 입력
출력 변수
예측된 중금속 고정화 효율(%)
예측값
고정화 효율 (%)
예측 오류(%)
예측 오류 (%)

도면13


시작
S100
토양 중금속 고정 예측을 위한 초기 데이터 세트 수집
S200
학습 알고리즘이 서로 다른 복수 개의 초기 머신 러닝 모델을 기반으로
초기 데이터 세트에 대한 토양 중금속 고정 예측 학습 각각 수행
S300
각각의 학습 결과에 대한 회귀 평가를 기반으로,
복수 개의 초기 머신 러닝 모델 중 토양 중금속 고정 예측을 위한
최적 머신 러닝 모델 설정
S400
초기 데이터 세트에 대하여, 입력 데이터의 특성(feature)의
중요도 및 입력 데이터의 특성간의 상관관계에 기반하여 필터링 수행
S500
필터링된 데이터 세트를 초기 데이터 세트에 추가하여
업데이트 데이터 세트 생성
S600
업데이트 데이터 세트를 이용하여 설정한 최적의 머신 러닝 모델 업데이트
종료