[서식 1] 공모계획서

「2020 연구데이터·AI 분석활용 경진대회」 공모계획서

프로젝트 제목	식물의 서식환경 공간 빅데이터 분석을 통한 ***'유용 식물 분포 가능성 예측지도'***제작

1. 프로젝트 제안 개요

○ 최근 세계 바이오산업 시장규모가 급속히 증가함에 따라 핵심을 이루는 생물 다양성의 경제적 가치도 함께 증가하고 있으며, 이에 따른 국가 간 생물 유전자원을 선점하려는 경쟁이 심화되고 있다. 따라서 우리나라도 **국가 생물자원 주권 확보에 대비**하고자 생물자원의 다양성 보전 및 관리, 지속적 이용 실현 등을 위한 광범위한 연구 활동 필요성이 요구되고 있다.

○ 국립생물자원관에서는 우리나라의 자생식물들을 대상으로 유용성 연구를 실시하고 있으며 이 가운데 **다방면으로의 산업화가 가능한 버드나무, 주목, 마, 은행나무의 유용성**에 대해 주목하고 있다.

○ 하지만, 국립생물자원관에서 제공하는 위 네 가지 식물의 생물자원분포도에는 다소 개략적인 식물의 분포위치 정보만 제공하여 상세한 서식지 분포와 서식에 적합한 토양과 지형 등의 환경정보를 제공하지 않고 있다. 따라서 이를 참고할 필요성이 있는 농림업 종사자, 국내 바이오 업체, 국가 생물연구 연구기관들에게는 이용의 한계가 있다.

○ 따라서 이번 프로젝트를 통해 **버드나무, 주목, 마, 은행나무가 성장하는데 적합한 토양, 지질, 지형, 기후 등 다양한 식물의 서식환경 공간 빅데이터 분석을 통하여 잠재 서식지를 예측**하고자 한다.

○ 전국자연환경조사 식물상 자료의 생물종 위치 데이터를 토대로 토양도, 지질도, 임상도, 수치지형도, 연평균 기온·강수의 환경변수 자료 데이터를 활용하여 **Logistic Regression과 MaxEnt 머신러닝 모델링 작업을 통해 잠재 서식지를 예측**, ***'유용 식물 분포 가능성 예측지도'***를 작성하여 본다.

○ ***'유용 식물 분포 가능성 예측지도'는*** 산업적 측면에서 국내 식물자원의 산업적, 경제적 가치를 높일 수 있을 것이고, 국가 간 생물자원 확보 경쟁에 있어서도 우리나라 생물자원의 주권을 보호할 수 있는 기초 연구 자료가 될 것으로 기대된다.

2. 제안 배경

○ 생물 다양성을 소재로 한 세계 바이오 의약품 시장이 2018년 기준으로 2,430억 달러로 전체 의약품 시장의 28%로 추정되며, 2024년 시장규모가 3,880억 달러에 달할 것으로 예상한다. 바이오산업의 시장규모가 커짐에 따라 핵심을 이루는 생물 다양성의 경제적 가치가 증가하며 생물 유전자원의 선점을 위한 경쟁이 심화되고 있다.

○ 식물 자원의 가치 있는 정보를 찾아내어 산업화 육성에 널리 활용하기 위해서는 범국가적으로 국가 생물자원의 다양성 보전 및 관리, 지속적 이용 실현, 생물 주권 보호 등을 위한 광범위한 연구 활동이 필요하다.

○ 국립생물자원관에서는 식물자원들 중 버드나무, 주목, 마, 은행나무 등을 산업화 가능성이 높은 대표적 유용 생물자원으로 선정하였다. 버드나무는 아스피린의 원료, 주목은 택솔로 항암제, 마는 스테로이드로 소염제, 은행나무는 혈액순환 개선제로 산업화가 가능한 생물자원이다.

○ 국립생물자원관의 생물지리정보시스템(공간분포정보)에서 제공하는 생물자원분포도에는 다소 개략적인 식물의 분포위치 정보만 제공하여 상세한 서식지 분포와 서식에 적합한 토양과 지형 등의 환경정보를 제공하지 않고 있다.

○ 따라서 산업적 가치가 있는 해당 식물을 재배하고 산업적으로 활용하고자 하는 농림업 종사자나 국내 바이오 업체, 더 나아가 국가 생물자원의 관리 및 이용을 위한 연구기관들 입장에서는 해당 식물이 자라기 적합한 환경과 그러한 환경적 조건을 갖춘 지역을 찾기에 어려움이 있다.

○ 이번 프로젝트를 통해서 산업적으로 유용한 식물인 버드나무, 주목, 마, 은행나무에 대하여 보다 상세한 서식 가능성이 높은 분포지역을 예측하고자 한다. 이러한 서식지 분포 가능성도 제작은 토양도, 임상도, 지질도, 지형도 등 식물 서식환경 공간 빅데이터를 기반으로 데이터 분석을 통하여 식물이 성장하는데 적합한 서식환경을 분석하여 예측하고자 한다.

3. 활용데이터 종류

데이터	포맷	출처
전국자연환경조사 자료 (육상식물, 식물상)	shp	에코뱅크 (https://www.nie-ecobank.kr)
토양도	shp	흙토람-농촌진흥청 (http://soil.rda.go.kr)
지질도	shp	지질정보서비스시스템 (https://mgeo.kigam.re.kr/)
임상도	shp	국가공간정보포털 (http://www.nsdi.go.kr)
수치지형도	shp	국토지리정보원 (https://www.ngii.go.kr)
연평균 기온, 연평균 강수량	csv	기상청-기상자료개방포털 (https://data.kma.go.kr)

4. 주요제안내용

○ 프로젝트 개요

- 생물종 위치자료 데이터를 바탕으로 토양도, 지질도, 임상도, 수치지형도, 연평균 기온·강수량의 지형, 기후 등 식물 서식 환경 공간 빅데이터 분석을 통하여 식물 서식환경과 관련된 환경변수를 추출한다. 환경변수들 중에서도 중요도가 높은 변수들만을 추출하여 Logistic Regression과 MaxEnt 모델의 머신러닝 작업을 통한 식물의 잠재 서식 가능지역을 예측한다.

○ 프로젝트 과정

[데이터 전처리 및 특성변환]

- 생물종 위치자료 데이터: QGIS를 활용하여 제2,3,4차 전국자연환경조사 식물상 데이터 중 분포 예측 대상인 주목, 마, 은행나무, 버드나무 데이터만을 추출해 낸다. 추출된 각 데이터를 식물 종 별로 합친 후 기존 벡터 형식을 레스터 형식으로 변환한다.
- 환경변수 자료 데이터: 민감도 분석을 통하여 각 환경변수가 식물 생장에 기여하는 정도를 평가한 후 중요도가 높은 변수들을 추려낸다.

[모델링을 통한 분포 예측]

- 전처리한 생물종 위치자료를 종속변수(y), 민감도 분석 이후 선택되어진 환경변수들을 독립변수(x)로 설정하여 train set과 test set을 구성하여 Logistic Regression 모델을 이용한 종 분포를 예측한다.

[모델간 비교]

- 생물 분포 위치와 환경과의 관계를 통계 해석하고 분포지역을 예측하는 알고리즘인 종분포모형(SDMs)을 활용한다. 종분포모형 중에서도 GUI 방식의 MaxEnt 모델을 이용하여 분석 대상 식물들의 분포를 예측한다. 이후 기존 Logistic Regression을 이용하여 예측한 모델과 MaxEnt를 이용한 예측 모델을 비교한다.

○ 프로젝트 결과

- 버드나무, 주목, 마, 은행나무가 자라나기에 적합한 환경 요소를 분석·모델링하여 식물 생장에 있어 적합한 토양과 지형 등의 환경정보를 확인하고 최종적으로 식물의 잠재 분포 가능지역을 예측하는 ***'유용 식물 분포 가능성 예측지도'***를 작성한다.

5. 기대효과

○ ***'유용 식물 분포 가능성 예측지도'***를 농림업 종사자와 생물자원을 산업적으로 활용하고자 하는 바이오 산업체에 제공하여 생물자원의 산업적, 경제적 가치를 높일 수 있을 것으로 기대한다.

○ 국내에 서식하고 있는 산업적 가치가 높은 생물자원의 분포를 알 수 있고, 기후변화에 따른 생물자원의 변화를 알 수 있어 우리나라 생물 주권을 효과적으로 보호하기 위한 수단으로 활용할 수 있다.

○ 바이오산업의 생물자원 활용 사업을 위한 기초자료도 활용될 수 있으며, 생물자원 정보관리체계 구축과 생태계 보호를 효율적으로 관리하는데 필요한 기초 연구 자료로도 활용될 수 있다.

6. KISTI제공 VM 오픈소스 라이브러리 및 버전 *(※ 분석에 활용할 툴 및 버전)*

※ 분석 최종결과물은 반드시 KISTI가 제공하는 VM에서 실행되어야 하며, 오픈소스 소프트웨어만 활용 가능함

※ 분석에 필요한 툴 종류와 버전을 명시(필요한 툴은 직접 설치가 원칙이며 필요시 설치를 지원할 수 있음)

※ KISTI에서 지원하는 VM 기본 환경

- H/W제공 : CPU: 2core, Memory : 16GB, Storage : 1TB - S/W제공 : python2x, 3x, Tensorflow 1x, 2x, Keras 등
-