



2025학년도 1학기 SW 캡스톤디자인 경진대회

유전자 데이터를 활용한 작물의 표현형 예측 모델 연구

오픈소스 URL : <https://github.com/shwoa/Capstone>

팀 명 알고리즘

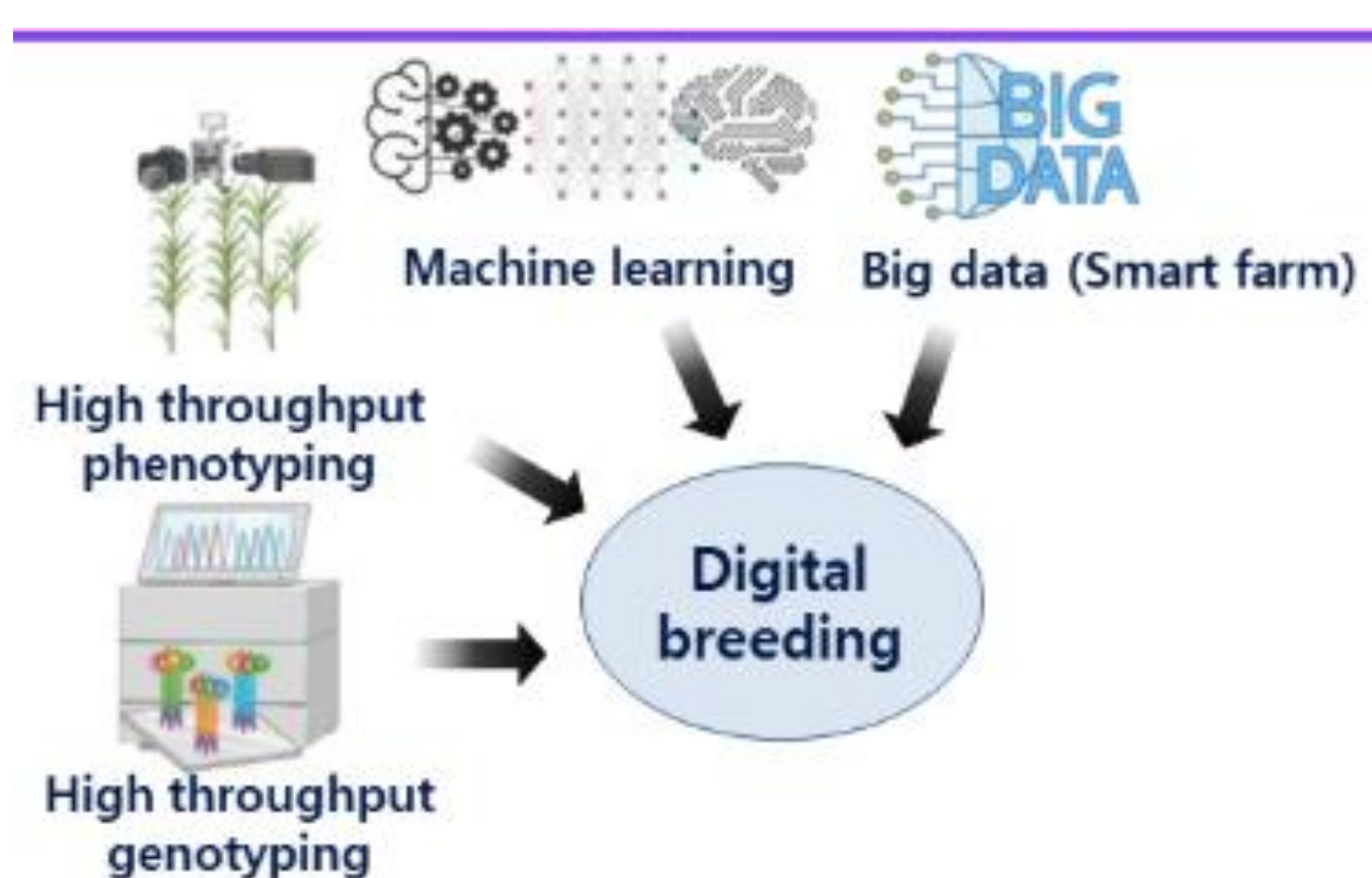
지도교수 고광신

팀 원 조진아(문헌정보학과, 4학년), 이하은(문헌정보학과, 4학년)

산업체 한국전자기술연구원

개발 동기 및 목적

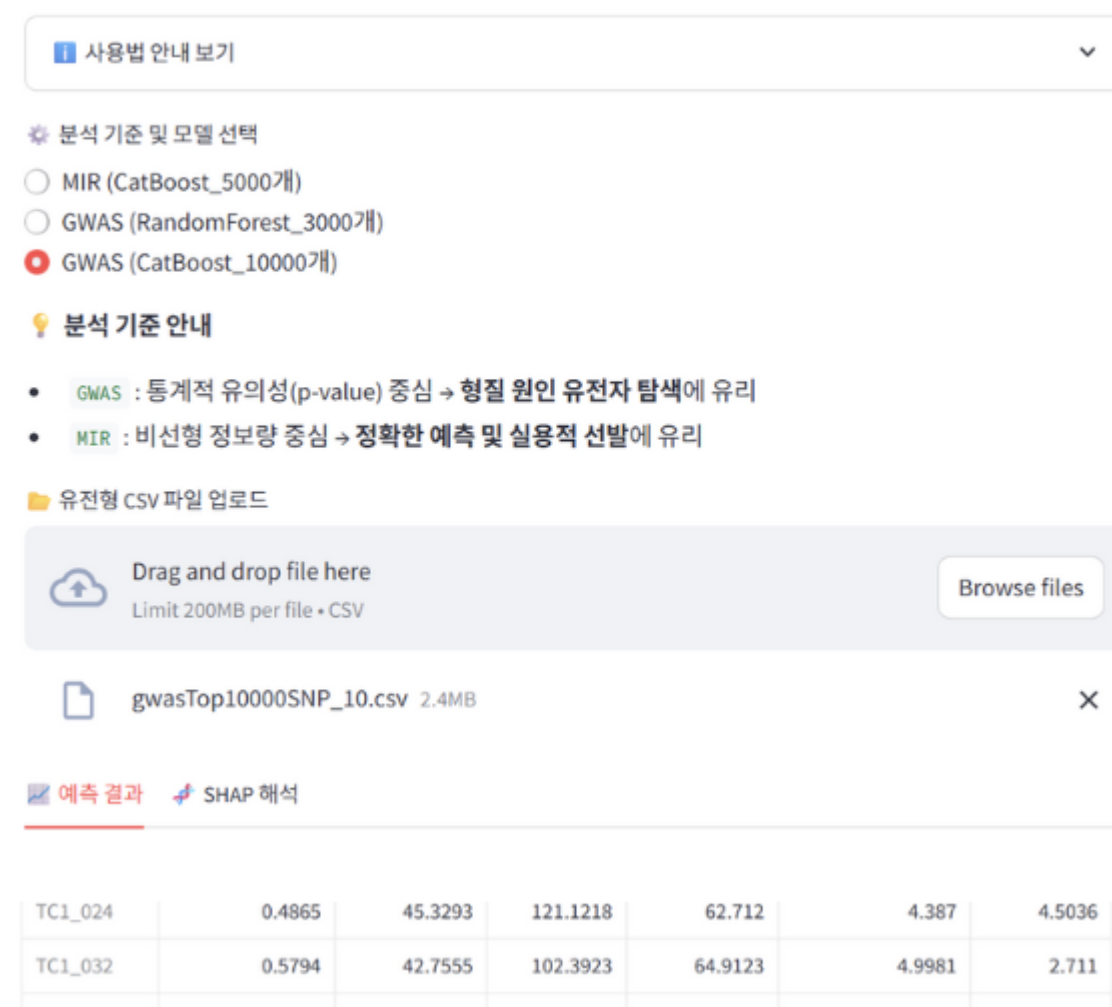
- 최근 기후 변화, 병충해 등으로 작물 수확 불안정성이 심화되었다.
- 기존 전통 육종 방식은 시행착오를 반복해야 하고 평균 5~10년 이상의 긴 개발 기간이 필요하며, 예측 기반 의사결정의 한계가 존재한다.
- 이에 유전형 정보를 기반으로 실험 없이 형질 예측을 하여 빠르고 효율적인 품종 개발이 가능한 디지털 육종이 대안으로 제시되었다.



- 이에 디지털 육종 기술을 활용하여, 토마토 품종의 데이터를 기반으로 AI 모델을 학습시키고, 누구나 활용 가능한 디지털 예측 웹사이트를 구현하고자 한다.

주요 기술

토마토 유전형 기반 표현형 예측기



원하는 분석 기준 및 AI 모델 선택

- 사용자에게 사전 지식 있다고 가정
- 추가로 상황별 분석 기준 안내 제공

유전형 CSV 파일 업로드

형질별 예측값 출력

결과 CSV로 다운로드 가능

“작물 형질을 예측하는 AI 모델을 웹 서비스로 제공하되, 사용자 수준에 맞춰 SNP 선별 기준과 해석 방법을 안내하고, 업로드된 유전형 데이터를 기반으로 형질 예측 결과를 출력 및 다운로드 할 수 있는 시스템을 구현한 기술”

- 데이터만으로 원하는 형질을 지닌 작물을 설계
- 형질이 좋은 작물을 미리 알아보고 골라낼 수 있도록 도와줌
- 누구나 쉽게 원하는 작물 형질을 예측 및 확인할 수 있게 함

개발 내용

[SNP 선별]

- 전체 51,000개 뉴 중 예측 성능 향상 위한 Top-N 선별했다.
- GWAS (p-value 기반 통계적 연관 분석)
- MIR (Mutual Information 기반 비선형 정보량 평가)
- SNP 수: 100, 500, 1000, 3000, 5000, 10000

[AI 모델 학습 및 성능 비교]

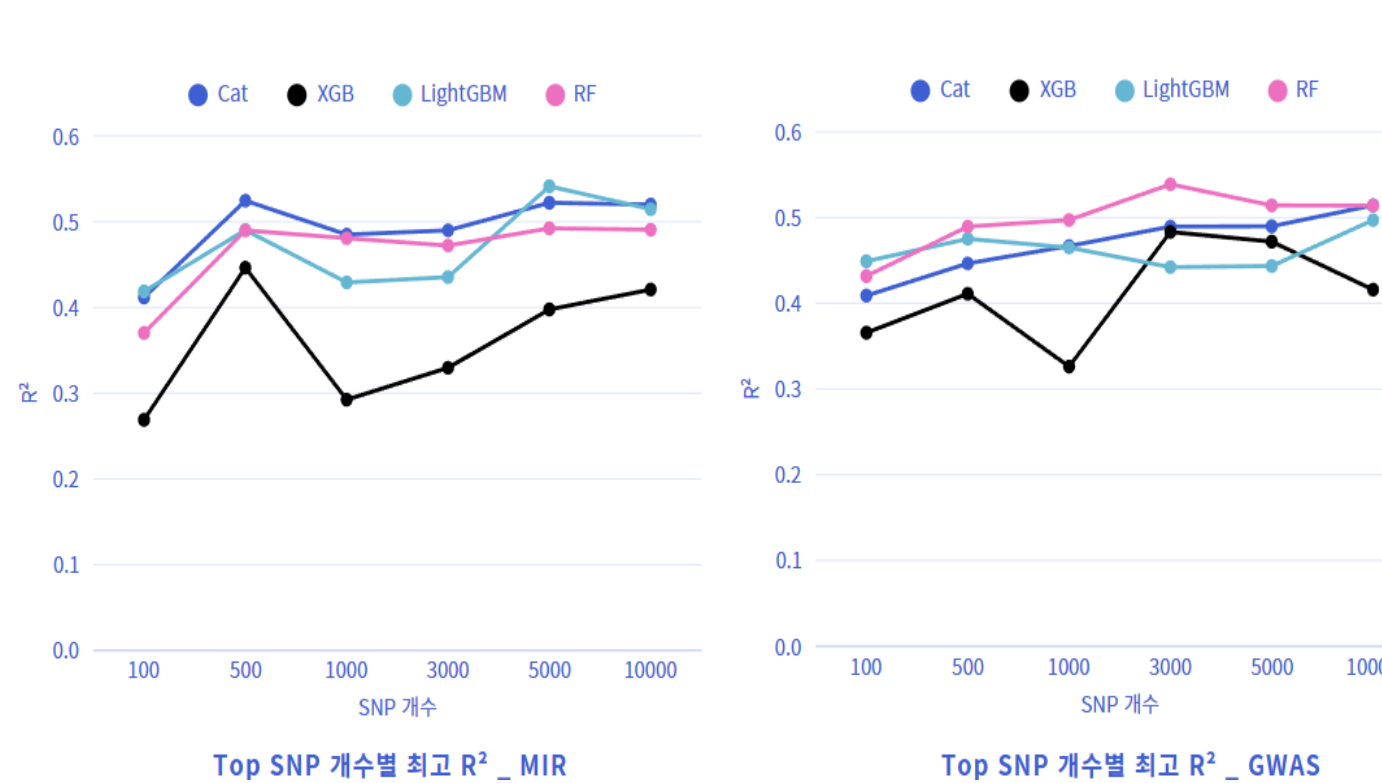
- AI 모델 종류: Tabnet, RF, XGB, LightGBM, CatBoost
- 평가 지표: 결정계수 R^2
- 실험 조건 : - 결측치 처리 방식 (최빈값 / 제거)
 - SNP 선별 방식 (GWAS / MIR)
 - SNP 개수 (100~10,000개)
- PyCaret 활용한 Automl로 성능 비교

[웹 서비스 구현]

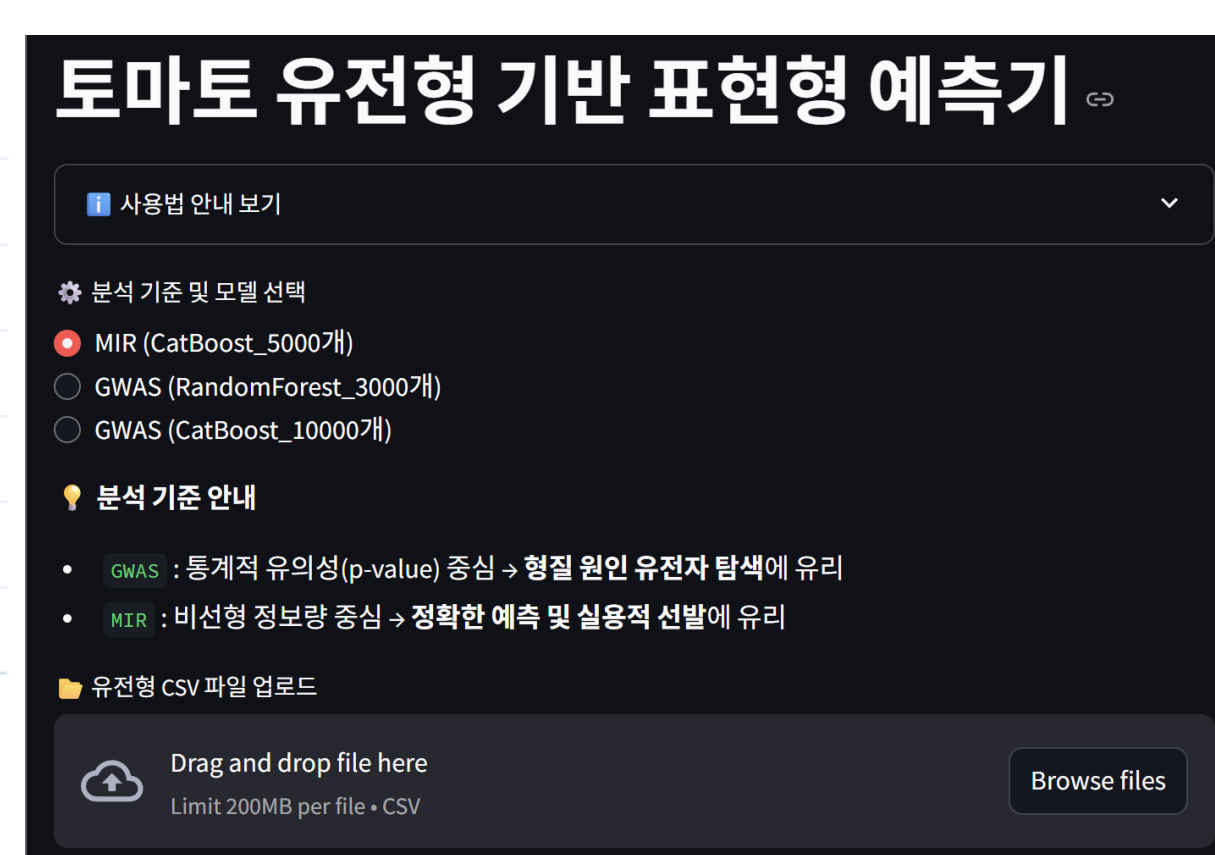
- AI 모델 실험 결과 선정된 우수 모델들을 사용하여 유전형 파일 업로드 하면 형질별 예측 결과를 출력하는 웹사이트 제작하였다.
- 형질별 예측 결과와 함께 형질별 SHAP 해석 시각화를 제공하였다.
- 결과 csv 파일로 다운로드 가능하도록 구현하였다.
- 분석 기준 및 모델 선택 가이드 함께 제공하였다.

결과 및 분석

[모델별 성능 및 웹 예측 시스템]



〈모델별 성능〉



〈웹 예측 시스템〉

[결과]

- CatBoost, XGBoost 등 머신러닝 모델을 활용해, 6개 표현형 예측 모델 구축
- CatBoost 모델이 가장 높은 예측 정확도(R^2)를 보임
- GWAS, MIR 기법으로 중요 SNP를 선별하여 모델 성능 향상
- SHAP 분석을 통해 중요 유전자 위치 시각화 및 해석 가능
- Streamlit 기반 웹 예측 시스템 개발 완료

[기대효과]

- 육종 기간 단축 및 실험 비용 절감
- 예측 결과에 대한 해석 가능성(XAI) 확보
- 농업 현장에서 쉽게 활용 가능한 도구로 기대
- 정밀 육종 및 디지털 육종 기술로 산업적 활용 가능성 높음



전북대학교
SW중심대학사업단