2024년도 공공기관 용역과제

AI개발 수행내역서

과제명	AI 기반 작물 최적화 추천 시스템
담당자	이　찬

2025년　　01 월　07 일

AI개발 수행내용

1. 사업과제 : 토양 및 환경 데이터 기반 최적 작물 추천 모델 개발

1. 개요 및 현황

2.1 추진배경 및 목적

○ 스마트팜 도입 증가로 인한 과학적 작물 선택 의사결정 지원 시스템 필요성 증대
○ 토양 영양분(질소, 인, 칼륨) 및 환경 조건에 따른 최적 작물 선정의 어려움
○ 농업 초보자들의 시행착오를 줄이고 수확률을 높이기 위한 AI기반 의사결정 지원 필요
○ 데이터 기반 작물 추천을 통한 농업 생산성 향상 및 자원 효율화 달성

2.2 과제 범위

과제구분		내용
AI 및 시각화	AI기반 수질예측모델 구현	원시 데이터 수집 및 데이터셋 구축
		데이터 전처리, 표준화, 상관관계 분석 (EDA도구 활용)
		Random Forest 모델 선정 및 학습
		Accuracy, precision 등 평가지표를 활용한 모델 성능 평가
		웹 프로토타입 구축
		예측모델 시각화
		streamlit을 통한 예측모델 웹기반 시스템 구축
		테스트

2.3 과제 추진 방법

1) 구축 대상 선정 기준

○ 데이터 접근성 및 활용성

- ■ 토양 성분과 환경 요인의 주요 독립변수 선정
- ■ 변수 간 상관관계 분석의 용이성
- ■ 모델의 해석 가능성 확보

○ 분석모델 개발 효율성

- ■ 영양소(질소, 인 칼륨)와 환경 요인(날씨, 습도, 산성도, 강수량) 그룹화를 통한 단계적 분석
- ■ 최소한의 독립변수로 작물 추천 정확도 확보

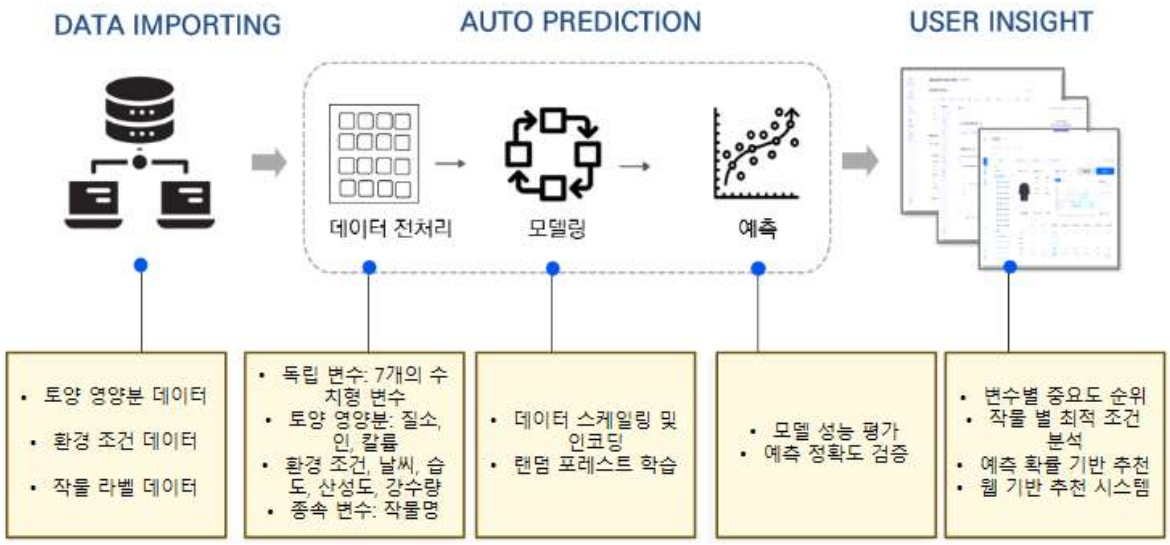
○ 분석모델의 기대 효과

- ■ 토양 조건별 최적 작물 추천을 통한 수확량 증대
- ■ 환경 조건에 따른 작물 선택 가이드라인 제시
- ■ 초보 농업인의 의사결정 지원을 통한 실패 위험 감소

2) AI 예측 분석모델 적용 대상

기능	수집 데이터	예측모델인자(독립변수)	AI 분석 대상
농작물 성장 요인 분석	- 토양 영양분 데이터 - 환경 조건 데이터	- 토양 성분: 질소, 인, 칼륨 - 환경 요인: 온도, 습도, pH, 강수량	- 주요 영향요인 식별 - 변수 간 상관관계 파악 - 작물 추천 결정 요인 분석

3) AI 분석모델 구축 프로세스



연구개발 주요 결과물

1. 데이터 수집

- 작물 추천 데이터 분석 리포트

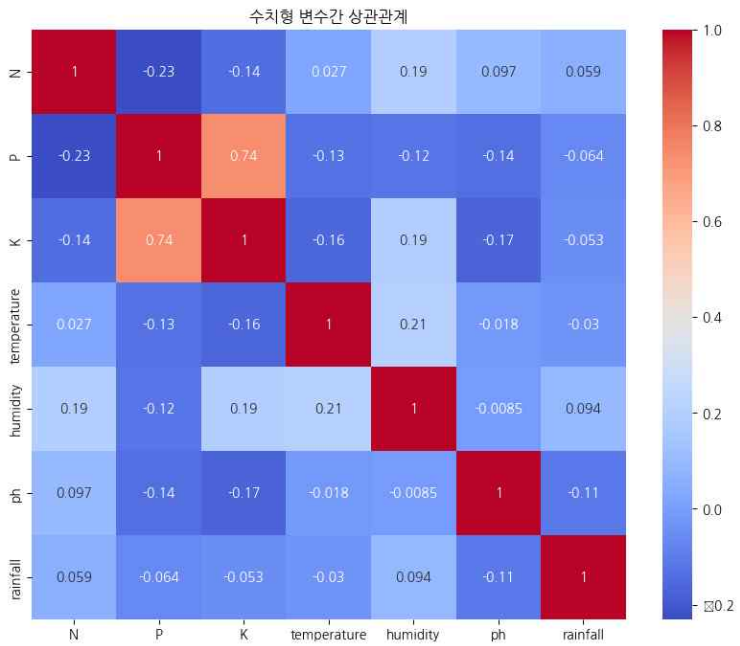
	A	B	C	D	E	F	G	H
1	N	P	K	temperature	humidity	ph	rainfall	label
2	90	42	43	20.87974	82.00274	6.502985	202.9355	rice
3	85	58	41	21.77046	80.31964	7.038096	226.6555	rice
4	60	55	44	23.00446	82.32076	7.840207	263.9642	rice
5	74	35	40	26.4911	80.15836	6.980401	242.864	rice
6	78	42	42	20.13017	81.60487	7.628473	262.7173	rice
7	69	37	42	23.05805	83.37012	7.073454	251.055	rice
8	69	55	38	22.70884	82.63941	5.700806	271.3249	rice
9	94	53	40	20.27774	82.89409	5.718627	241.9742	rice
10	89	54	38	24.51588	83.53522	6.685346	230.4462	rice
11	68	58	38	23.22397	83.03323	6.336254	221.2092	rice
12	91	53	40	26.52724	81.41754	5.386168	264.6149	rice
13	90	46	42	23.97898	81.45062	7.502834	250.0832	rice
14	78	58	44	26.8008	80.88685	5.108682	284.4365	rice
15	93	56	36	24.01498	82.05687	6.984354	185.2773	rice
16	94	50	37	25.66585	80.66385	6.94802	209.587	rice
17	60	48	39	24.28209	80.30026	7.042299	231.0863	rice
18	85	38	41	21.58712	82.78837	6.249051	276.6552	rice
19	91	35	39	23.79392	80.41818	6.97086	206.2612	rice
20	77	38	36	21.86525	80.1923	5.953933	224.555	rice
21	88	35	40	23.57944	83.5876	5.853932	291.2987	rice
22	89	45	36	21.32504	80.47476	6.442475	185.4975	rice
23	76	40	43	25.15746	83.11713	5.070176	231.3843	rice

(출처: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset)

- 데이터 규모: 2,200개 측정 데이터

2. 데이터 분석

2.1 변수 간 상관관계(Heatmap)



- 인과 칼륨 사이에 강한 양의 상관관계(0.736)
- 실제 농업에서 두 영양소가 비료 등을 통해 함께 공급 됨이 원인
- 대다수의 변수들은 서로는 약한 상관관계를 보임(각 변수들이 비교적 독립적이다)
- 질소와 인 사이에 약한 음의 상관관계(-0.231), 질소와 칼륨 사이에도 약한 음의 상관관계(-0.141)
- 칼륨이 너무 많으면 이온 길항작용/삼투압 균형/뿌리 발달 저하와 같은 이유로 인해 질소 흡수가 줄어듦. 실제 영양분 분배 패턴과 현 데이터셋은 동일한 패턴을 가짐

2.2 탐색적 데이터 분석

Overview | Alerts 11 | Reproduction

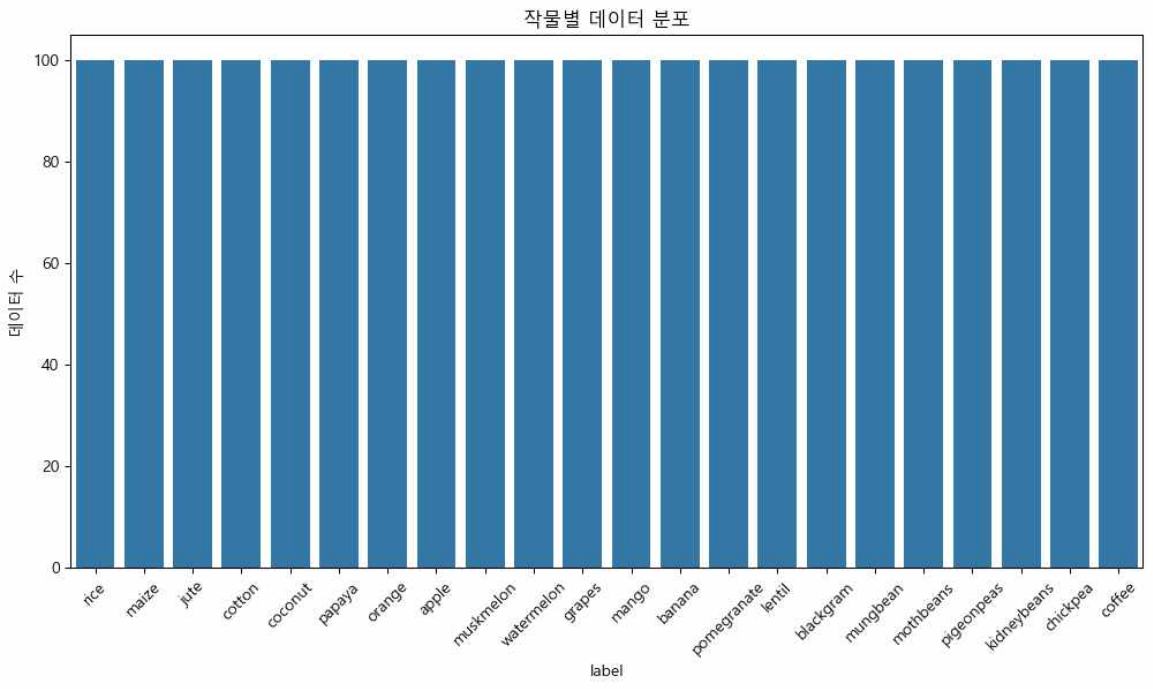
Dataset statistics

Number of variables	8
Number of observations	2200
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	137.6 KiB
Average record size in memory	64.1 B

Variable types

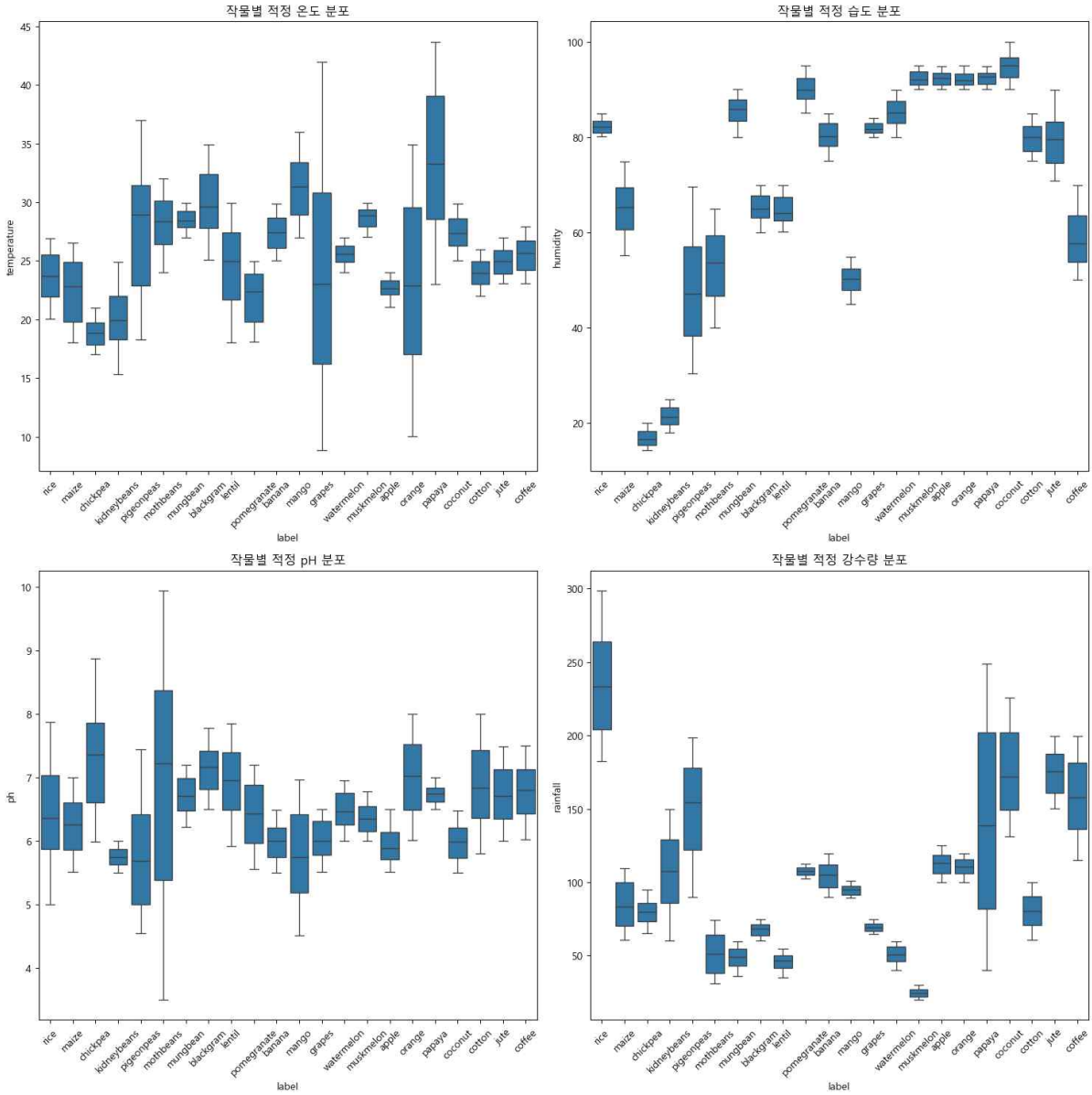
Numeric	7
Categorical	1

○ 작물별 데이터 분포



■ 총 22개의 항목에 각 100개의 데이터 수를 가짐

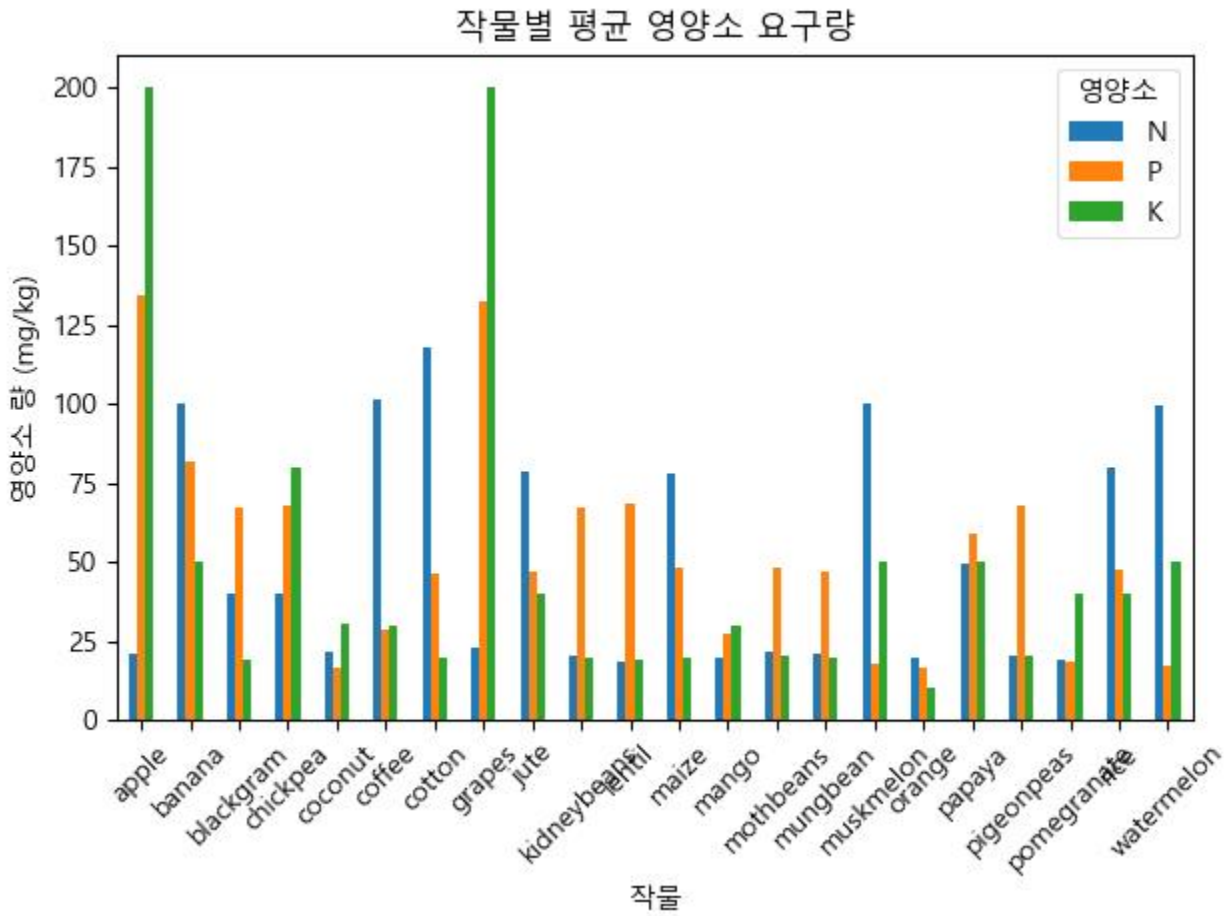
○ 변수 별 데이터 분포



1. 온도 분포
   - ■ 파파야가 가장 높은 온도 선호 / 병아리콩이 가장 낮은 온도 선호
   - ■ 대부분 작물이 20~30C범위에 분포
2. 습도 분포
   - ■ 코코넛, 파파야, 망고 등 열대과일이 높은 습도 선호 / 병아리콩이 가장 낮은 습도 선호
   - ■ 작물 간 습도 요구량 차이가 매우 큼
3. pH 분포
   - ■ 대부분의 작물이 pH6-7의 범위에서 재배 가능
4. 강수량 분포
   - ■ 쌀이 가장 높은 강수량 요구 / 멜론류가 가장 낮은 강수량 요구

분석 의의: 작물 별 최적 재배 조건 파악

○ 작물별 영양소 요구량



○ 특이사항

- 질소, 인, 칼륨 요구량이 각자 매우 높음
- 포도와 사과는 비슷한 영양소 패턴을 보임
- 콩류는 대체로 비슷한 영양소 요구량을 보임

○ 데이터 전처리

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
5	69	37	42	23.058049	83.370118	7.073454	251.055000	rice
6	69	55	38	22.708838	82.639414	5.700806	271.324860	rice
7	94	53	40	20.277744	82.894086	5.718627	241.974195	rice
8	89	54	38	24.515881	83.535216	6.685346	230.446236	rice
9	68	58	38	23.223974	83.033227	6.336254	221.209196	rice

apple (사과)/banana (바나나)/blackgram (우라드콩)/chickpea (병아리콩)/coconut (코코넛)/coffee (커피)/cotton (목화)/grapes (포도)/jute (황마)/kidneybeans (강낭콩)/lentil (렌틸콩)/maize (옥수수)/mango (망고)/mothbeans (모스빈 – 인도산 강낭콩)/mungbean (녹두)/ muskmelon (머스크멜론)/orange (오렌지)/papaya (파파야)/pigeonpeas (비둘기콩)/pomegranate (석류)/rice (쌀)/watermelon (수박)

3. 데이터 학습 및 모델정의

3.1 모델정의: 랜덤 포레스트

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import classification_report, confusion_matrix
import warnings
warnings.filterwarnings('ignore')

# 한글 폰트 설정
plt.rc('font', family='Malgun Gothic')
```

3.2 모델학습 및 학습 시각화

○ 모델 학습

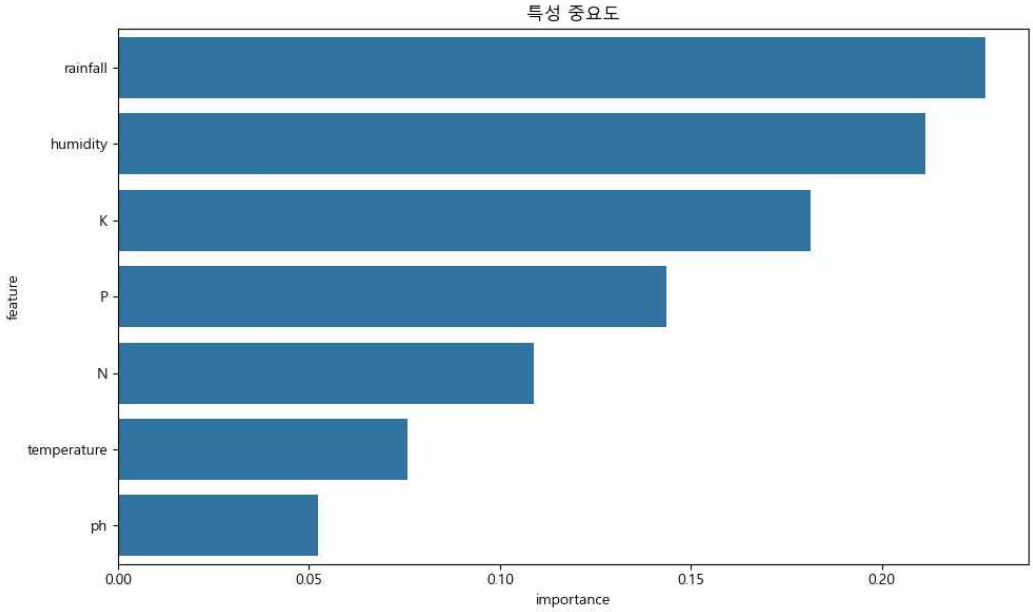
```
# 여덟 번째 셀 - 랜덤 포레스트 모델 학습
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train_scaled, y_train)

# 모델 성능 평가
train_score = rf_model.score(X_train_scaled, y_train)
test_score = rf_model.score(X_test_scaled, y_test)

print("학습 데이터 정확도: {:.2%}".format(train_score))
print("테스트 데이터 정확도: {:.2%}".format(test_score))
```

```
학습 데이터 정확도: 100.00%
테스트 데이터 정확도: 99.32%
```

○ 특성 중요도 시각화

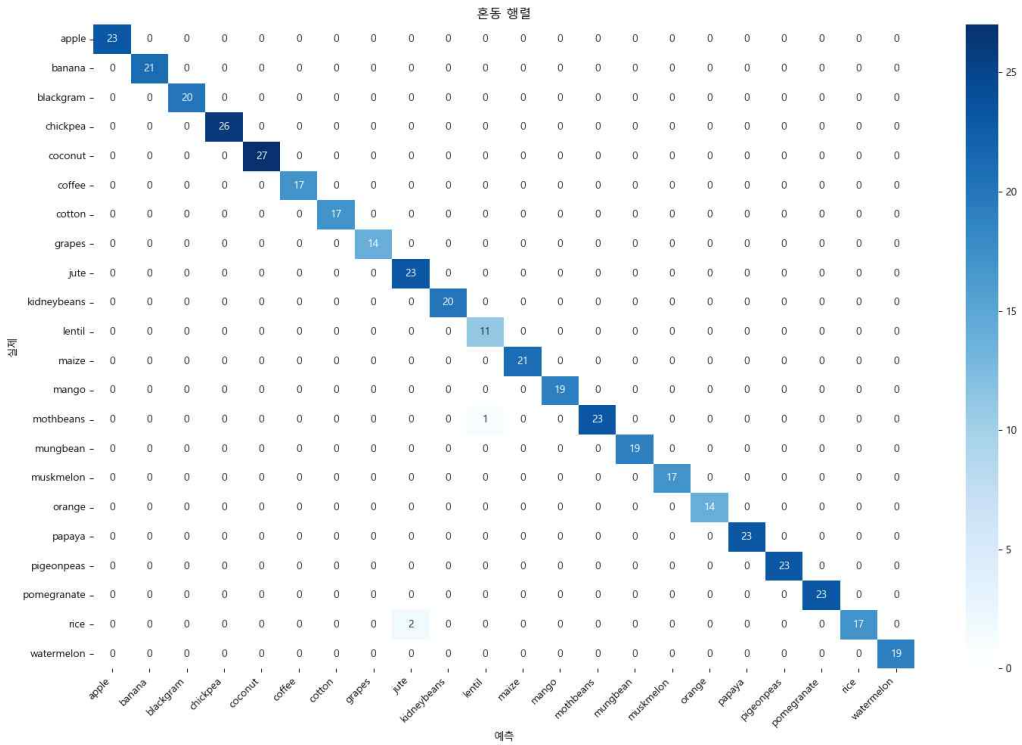


■ 환경 요인이 가장 중요함을 시사

■ 토양 영양분의 계층적 중요도: 칼륨>인>질소 순으로 중요. 세 영양소가 전체 영향력의 43.4% 차지

■ 온도와 산성도의 낮은 중요도

○ 혼동 행렬 시각화



■ 모든 작물에 대해 고른 예측 성능을 보임

○ 상위 추천 작물

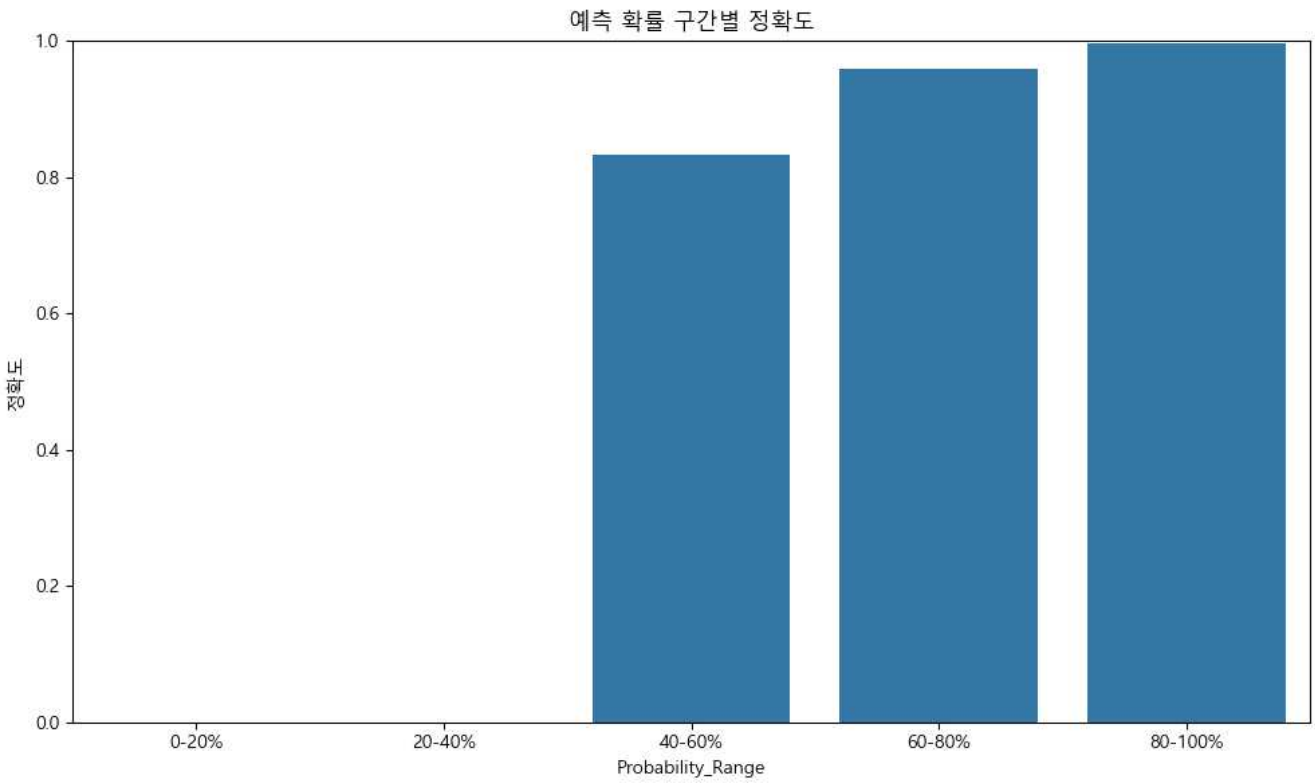
질소 함량 = 90 / 인 함량 = 40 / 칼륨 함량 = 40 / 온도: 20도 / 습도= 80 / 산성도 = 7 / 강수량 = 200

```
추천 작물: rice

상위 3개 추천 작물 및 확률:
rice: 75.00%
jute: 25.00%
apple: 0.00%
```

3.3 모델 예측

○ 예측 확률 정확도



○ 모델 성능 평가

```
print("모델 성능 평가:")
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
```

✓ 0.0s

```
모델 성능 평가:
Accuracy: 0.9932
Precision: 0.9937
```

■ 데이터의 특성

작물마다 필요한 생육 조건이 명확히 구분됨
질소, 인, 칼륨, 온도, 습도 등이 작물별로 뚜렷한 패턴을 보임
실제 농업 지식과 일치하는 패턴

■ 성능 평가

훈련 데이터: 100%
테스트 데이터: 99.32%
두 데이터셋의 성능 차이가 매우 작음 (0.68%)

■ 혼동 행렬 분석

대부분의 작물이 정확히 분류됨
발생한 오분류도 비슷한 특성의 작물 간에 발생
무작위한 오류가 아닌 패턴이 있는 오류

■ 예측 확률 분포

대부분의 예측이 높은 신뢰도를 보임
낮은 확률의 불확실한 예측이 거의 없음

따라서 이 높은 정확도는 과적합이라기보다는, 명확한 패턴을 가진 데이터를 잘 학습한 결과로 판단

1. 프로토타이핑(화면)

4.1 모델 예측

○ 스마트 작물 추천 시스템