과제구분	기관고유	수행시기		전반기	
연구과제 및 세부과제		연구분야	수행기간	연구실	책임자
데이터 기반 인공지능 농업적 활용 연구		농업정보	'22~'27	원예연구과	김혜형
머신러닝 알고리즘을 활용한 과채류 수확량 예측 모델 개발		농업정보	'22~'25	원예연구과	**김혜형**
색인용어	스마트팜, 빅데이터, 과채류, 머신러닝, 모델				

ABSTRACT

The utilization of smart farm environmental and growth data for yield prediction is increasingly recognized as a pivotal technology in facilitating data-driven agricultural decision-making. In this study, we aimed to develop a model for predicting fruiting vegetables yield using smart farm data and implement it as a service integrated with an information system. To establish a systematic framework for managing and analyzing smart farm data, we initially developed a data standard dictionary and designed a data pipeline structure comprising a data lake, data warehouse, and data mart, thereby creating robust environment for data analysis. For the development of the yield prediction model, we constructed an analytical dataset that integrates smart farm environmental and growth data.

The cucumber yield prediction model employed machine learning algorithms with input variables such as cultivation week, internal temperature, internal relative humidity, internal carbon dioxide concentration, and external solar radiation. For the cherry tomato yield prediction model, a total of 38 analytical variables based on growth and environmental data were constructed, and key variables selected through correlation analysis. Upon comparing various machine learning algorithms, the cucumber model achieved a mean absolute error (MAE) of 0.448 and a coefficient of determination (R^2) of 0.714 using the XGBoost algorithm, while the cherry tomato model achieved a MAE of 8.04 and an R^2 of 0.88 using the Gradient Boosting algorithm.

The developed prediction models were integrated with the Gyeonggi-do Smart Farm Data Utilization Service (GSDUS), automating processes such as environmental data collection, growth data upload, data preprocessing, and generation of derived variables. Prediction results are visualized as weekly graphs illustrating trends in

yield changes throughout the cultivation period. The findings of this study underscore the potential for developing a data-driven smart agriculture decision support service that predicts fruiting vegetables yield based on smart farm data for practical application.

Key words: Smart farm, Big data, Fruiting vegetables, Machine learning, Model

1. 연구목표

최근 스마트팜 보급 확대와 함께 재배 환경 및 생육 데이터를 활용한 데이터 기반 농업기술 개발의 중요성이 증가하고 있다. 스마트팜에서는 다양한 센서와 시스템을 통해 환경 데이터가 지속적으로 수집되고 있으나 장비 간 호환성 문제, 데이터 구조의 표준화 부족 및 분석 기반의 미흡 등으로 인해 실제 농업 현장에서의 활용도는 제한적인 실정이다(노 등, 2020; 김 등, 2022). 또한 시설원예 작물의 생산량은 온실 환경, 생육 상태 및 재배 관리 요인의 복합적인 영향을 받기 때문에 이를 반영한 생산량 예측 기술 개발이 필요한 것으로 보고되고 있다(Lin 등, 2019).

스마트팜에서 수집되는 환경 데이터는 작물 생육 상태 분석과 생산량 예측을 위한 기초 자료로 활용될 수 있으며, 토마토 재배 데이터를 이용한 분석 연구에서도 작물 생육 지표와 온실 환경 조건이 수확량과 높은 상관관계를 나타내는 것으로 보고된 바 있다(김 등, 2023; 노 등, 2020; 한 등, 2023). 최근에는 환경 데이터와 생육 데이터를 활용한 머신러닝 기반 작물 생산량 예측 연구가 활발히 수행되고 있으며 ANN, Random Forest, SVR 등 다양한 머신러닝 기법이 작물 생산량 예측에 적용되어 높은 예측 정확도를 보이는 것으로 보고되고 있다(Odah 등, 2025; Mancer 등, 2024). 또한 이러한 분석 기술은 작물 생육 상태와 생산량을 예측하여 재배 관리 및 생산 계획 수립을 지원하는 스마트농업 의사결정 지원 기술로 활용될 수 있다(Gong 등, 2021).

따라서 본 연구에서는 스마트팜에서 수집되는 환경 및 생육 데이터를 활용하여 시설원예 작물(오이, 방울토마토)의 생산량 예측 모델을 개발하고, 데이터 기반 분석 환경 구축과 머신러닝 기반 생산량 예측 모델 개발을 통해 데이터 기반 스마트농업 의사결정 지원 기술을 확보하고자 하였다. 또한 개발된 생산량 예측 모델을 정보시스템과 연계하여 실제 농업 현장에서 활용 가능한 생산량 예측 서비스 형태로 구현하고자 하였다.

2. 재료 및 방법

<시험1> 스마트팜 데이터 기반 분석 환경 구축

스마트팜 데이터는 다양한 장비와 시스템에서 서로 다른 형식으로 생성되기 때문에 데이터 표준화와 데이터 관리 체계 구축이 데이터 기반 농업 연구의 중요한 선행 단계로 제시되고 있다. 본 연구에서는 데이터 표준사전 구축과 데이터 파이프라인

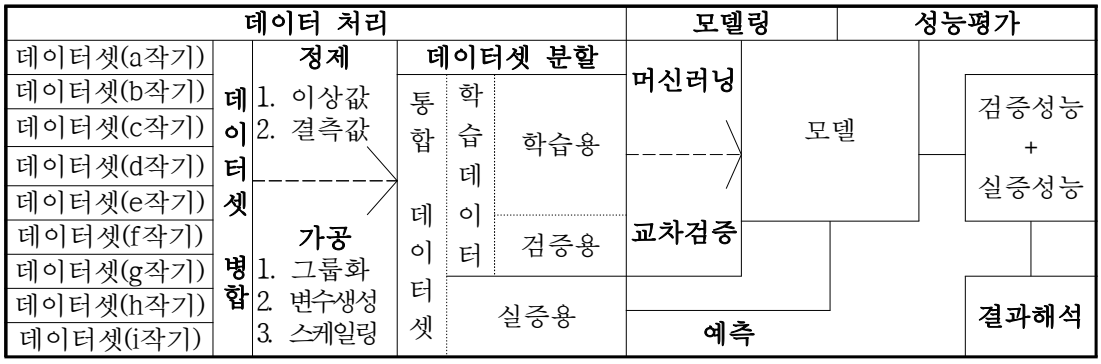
설계를 통해 스마트팜 데이터를 분석 가능한 형태로 통합 관리할 수 있는 데이터 인프라를 구축하였다.

데이터 표준화를 위해 단어사전, 용어사전, 도메인사전, 코드사전으로 구성된 데이터 표준사전을 작성하였다. 또한 공공기관 데이터베이스 표준화 지침을 참고하여 데이터베이스 관리 항목을 정의한 후 데이터베이스 정의서를 작성하였다. 스마트팜 센서 데이터를 분석에 활용하기 위해 데이터 수집, 정제 및 가공 과정을 포함하는 데이터 파이프라인을 구축하였다. 데이터 처리 과정은 추출·변환·적재(ETL) 구조로 설계하였으며, 데이터는 데이터 레이크(Data Lake), 데이터 웨어하우스(Data Warehouse), 데이터 마트(Data Mart) 구조로 관리되도록 하였다. 이를 통해 최종적으로 분석에 활용 가능한 분석용 데이터셋을 생성할 수 있도록 하였다.

<시험2> 머신러닝 기반 생산량 예측 모델 개발

스마트팜에서 수집되는 환경 및 생육 데이터를 활용하여 시설원예 작물(오이, 방울토마토)의 생산량 예측 모델을 개발하였다. 분석 절차는 데이터 수집, 데이터 전처리, 탐색적 데이터 분석(EDA), 머신러닝 모델 구축 및 성능 평가 단계로 구성하여 수행하였다(표 1).

표 1. 수확량 예측 모델 개발을 위한 데이터 분석 체계도



분석 데이터는 스마트팜 환경센서를 통해 수집된 환경데이터와 재배 현장에서 조사된 생육데이터를 활용하였다. 환경데이터는 내부온도, 내부상대습도, 내부 이산화탄소 농도, 외부 일사량 등 스마트팜 센서를 통해 수집되는 자료를 사용하였으며, 생육데이터는 작물 생육 조사 결과를 활용하였다. 또한 일부 환경 변수의 결측 데이터는 기상자료개방포털의 농업기상관측(AAOS) 자료를 활용하여 보완하였다.

오이 수확량 예측 모델 개발을 위해 경기도 오이 재배 16농가를 대상으로 2018년부터 2022년까지 수집된 작기 단위 데이터를 65건을 활용하였으며, 방울토마토 생산량 예측 모델 개발을 위해 경기도 방울토마토 재배 12농가를 대상으로 2017년부터 2023년까지 수집된 작기 단위 데이터 74건을 활용하였다(표 2).

표 2. 생산량 예측 모델 개발을 위한 데이터셋 구성

작목	농가수	데이터 수집기간	데이터셋
오이	16농가	2018~2022	65
방울토마토	12농가	2017~2023	74

스마트팜 센서 데이터는 시간 단위로 수집되므로 분석에 활용하기 위해 재배 주차 기준으로 집계하여 분석용 데이터셋을 구축하였다. 환경데이터는 평균값 및 누적값을 계산하여 변수로 생성하였으며, 결측치 제거 및 단위 변환 등 데이터 정제 과정을 수행하였다. 또한 작물 생육 데이터와 환경 데이터를 재배 주차 기준으로 통합하여 분석용 데이터셋을 구성하였다.

스마트팜 환경 및 생육 데이터 분석과 생산량 예측 모델 구축은 Python 기반 분석 환경에서 수행하였다. 독립변수는 MinMaxScaler를 이용하여 정규화하였으며, 학습용 데이터와 검증용 데이터로 분할하여 분석하였다. 환경 변수와 생산량 간의 관계를 파악하기 위해 기초 통계 분석 및 시각화를 수행하였으며, 변수 간 상관관계 분석을 통해 생산량과 관련성이 높은 변수를 탐색하고 모델 입력 변수 후보를 선정하였다.

생산량 예측 모델 개발을 위해 회귀 기반 알고리즘(LinearRegression, Ridge, Lasso 등)과 앙상블 기반 알고리즘(RandomForest, XGBoost, Gradient Boosting 등)을 적용하여 생산량 예측 성능을 비교하였다. 모델 성능평가는 평균절대오차(MAE)와 결정계수(R^2)를 활용하였으며, 모델 학습과 검증 과정은 학습 데이터와 테스트 데이터를 구분하여 수행하고 교차 검증을 통해 모델 성능을 비교하였다.

<시험3> 생산량 예측 모델 시스템 구현

<시험2>에서 개발된 오이 및 방울토마토 생산량 예측 모델을 자체 운영 중인 정보시스템인 '경기도스마트팜데이터활용서비스(GSDUS)'에 구현하고 서비스하고자 하였다. Python 기반으로 개발된 예측 모델이 시스템에서 실행될 수 있도록 모델 실행 환경을 구성하였으며, 독립변수로 활용된 환경데이터 및 생육데이터가 시스템에 연계되도록 구현하였다. 데이터와 예측 모델이 연계된 후 재배 주차별 생산량 예측 결과가 자동으로 생성되도록 하였으며, 분석 결과는 사용자 인터페이스를 통해 그래프로 시각화하여 확인할 수 있도록 하였다.

3. 결과 및 고찰

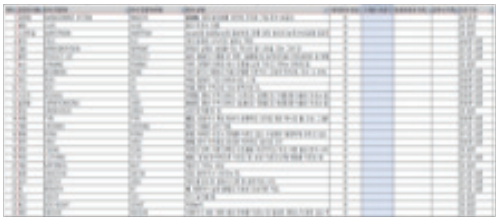
<시험1> 스마트팜 데이터 기반 분석 환경 구축

스마트팜 데이터의 체계적 관리와 분석 활용을 위해 데이터 표준사전 구축과 데이터 파이프라인 설계를 수행하였다. 먼저 스마트팜 데이터 기반 분석 환경을 구축하기 위해 자체 운영 중인 정보시스템 '경기도스마트팜데이터활용서비스(GSDUS)'의 데이터베이스 표준화 및 구조 정비를 수행하였다.

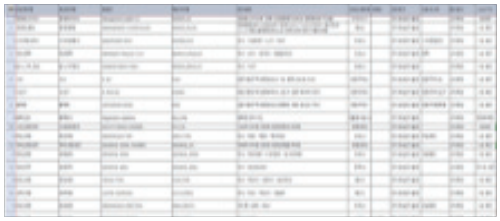
'시설원예 분야 스마트팜 수집 데이터 규격'을 참고하여 데이터 항목을 정의하고 '공공기관 데이터베이스 표준화 지침'에 따라 데이터 표준사전을 작성하였다. 데이터 항목 간 용어 불일치를 최소화하고 데이터 통합 및 분석 활용성을 높이기 위해 단어, 용어, 도메인, 코드 사전으로 구성된 데이터 표준사전을 구축하였으며 그 결과 단어 162개, 용어 157개, 도메인 37개, 코드 104개로 구성된 데이터 표준사전을 구축하였다(표 3, 그림 1). 또한 데이터베이스 구조를 체계적으로 관리하기 위해 데이터베이스 정의서, 엔터티 정의서, 속성 정의서 등 데이터베이스 관리 문서를 작성하여 데이터 구조를 정립하였다(표 4, 그림 2). 이러한 데이터 표준화와 관리체계 구축의 필요성은 디지털 농업 촉진을 위한 농업 빅데이터 및 인공지능 활용 기술 연구에서도 제시된 바 있으며, 데이터 수집 규격 정의와 데이터 표준 관리 체계 구축의 중요성이 보고된 바 있다(노 등, 2020; R. Solanki, 2025).

표 3. 데이터 표준사전 관리 항목

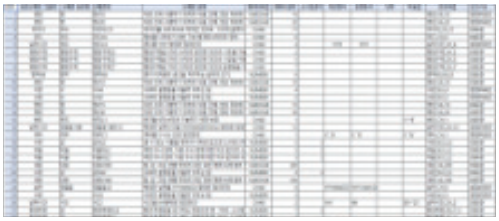
구 분	관리 항목	결과
단어사전	표준단어명, 단어 영문명, 단어 영문약어명, 단어 설명, 형식단어 여부, 도메인 분류명, 이음동의어 목록, 금칙어 목록	162개 단어
용어사전	표준용어명, 영문명, 영문약어명, 용어설명, 표준도메인명, 허용값, 관리부서명, 표준코드명, 업무분야	157개 용어
도메인사전	표준도메인, 그룹명, 도메인분류명, 도메인명, 도메인 설명, 데이터타입, 데이터길이, 소수점 길이, 저장형식, 표현형식, 단위, 허용값	37개 도메인
코드사전	관리부서명, 한글코드명, 영문코드명, 코드설명, 데이터타입, 데이터길이, 코드값, 코드값 의미	104개 코드



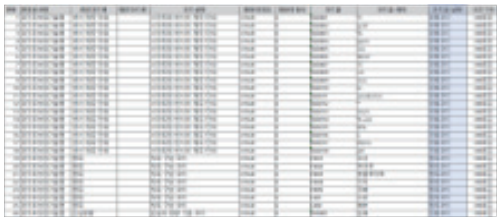
[표준 단어사전]



[표준 용어사전]



[표준 도메인사전]



[표준 코드사전]

그림 1. 항목별 데이터 표준사전

표 4. 데이터베이스 산출물 표준 관리 항목

구 분	주요 항목	결과
데이터베이스 정의서	기관명, 부서명, 관련법령, 한글 DB명, 영문 DB명, 구축일자, DB 설명, 업무분류체계, DBMS 정보, 운영체제정보, DB 형태	4개 D/B 정의
논리데이터모델 다이어그램	한글 DB명, 설명	관계 논리 구조화
엔터티정의서	한글DB명, 엔터티명, 엔터티, 설명, 주식별자, 수퍼타입, 엔터티명, 관련 엔터티명, 테이블 설명, 발생주기	21개 개체 정의
애트리뷰트 정의서	엔터티명, 속성명, 속성유형, 필수입력여부, 식별자 여부, 참조, 엔터티명, 참조, 속성명, 속성설명	289개 속성 정의

[기초 데이터베이스 정의서]

[데이터레이크 정의서]

[데이터웨어하우스 정의서]

[데이터마트 정의서]

그림 2. 항목별 데이터베이스 정의서

또한 김 등(2022)은 스마트농업 확산을 위해 스마트팜 장비와 데이터 구조의 표준화가 중요하다고 보고하였다. 이에 본 연구에서는 스마트팜 센서 데이터를 효율적으로 관리하고 분석에 활용하기 위해 데이터 레이크(Data Lake), 데이터 웨어하우스(Data Warehouse), 데이터 마트(Data Mart) 구조의 데이터 파이프라인을 구축하였다. 이를 통해 수집 데이터의 저장, 정제 및 분석 활용 과정을 단계적으로 수행할 수 있도록 하였으며 데이터 정제 및 변환을 위한 ETL 프로세스를 구현하였다(그림 3). 이와 같은 데이터 표준화와 데이터 파이프라인 구축을 통해 스마트팜 데이터의 체계적인 관리와 분석 기반 연구 수행을 위한 데이터 인프라를 구축하였다.

김 등(2023)에 따르면 스마트팜 환경 데이터의 체계적인 수집과 관례 체계 구축은 작물 생육 분석 및 생산량 예측 연구를 수행하기 위한 필수적인 기반으로 보고된 바 있으며, 센서 기반 환경 데이터의 표준화와 데이터 관리 체계 구축이 스마트농업 연구의 중요한 요소로 제시되고 있다.

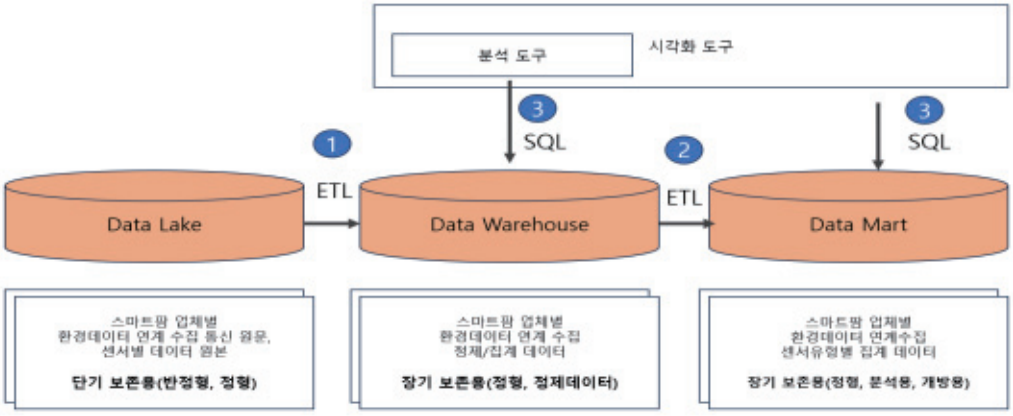


그림 3. 스마트팜 데이터 분석을 위한 데이터 파이프라인 구조

<시험2> 머신러닝 기반 생산량 예측 모델 개발

스마트팜 환경 및 생육 데이터를 활용하여 시설원예 작물의 생산량 예측 모델을 개발하기 위한 분석용 데이터셋을 구축하였다. 오이 수확량 예측 모델 개발을 위해 2018년부터 2022년까지 수집된 작기 단위 데이터 65건을 활용하였으며, 이 중 결측 데이터가 많거나 필요한 변수 항목이 모두 포함되지 않은 데이터를 제외한 분석 가능한 43건의 데이터셋을 병합·정제·가공하여 최종 분석용 통합 데이터셋을 생성하였다. 방울토마토 생산량 예측 모델 개발을 위해 2017년부터 2023년까지 수집된 작기 단위 데이터 74건 중 16주 이상 방울토마토를 재배한 47건의 데이터셋을 활용하였다.

스마트팜 재배 데이터를 활용한 연구에서도 작물 생육 지표와 온실 환경 조건이 수확량과 밀접한 관계를 나타내는 것으로 보고된 바 있으며, 환경 데이터와 생육 데이터를 통합한 분석을 통해 생산량 예측 연구가 수행되고 있다(노 등, 2020; 한 등, 2023).

환경 데이터는 내부온도, 내부상대습도, 내부 이산화탄소 농도, 외부 일사량 등 스마트팜 센서를 통해 수집된 자료를 활용하였으며, 일부 환경 변수의 결측 데이터는 기상자료개방포털의 농업기상관측 자료를 활용하여 보완하였다. 또한 1시간 단위로 수집되는 스마트팜 센서 데이터는 재배 주차 기준의 1주 단위 데이터로 변환하여 분석용 변수로 활용하였다.

오이 수확량 예측 모델의 경우 환경 센서를 통해 자동 수집될 수 있고 누락된 데이터를 대체할 수 있는 환경 변수를 선정하였다. 독립 변수는 정식 후 재배 주차, 내부온도, 내부상대습도, 내부 이산화탄소 농도, 외부 일사량 등 5개 변수로 구성하였다(표 5). 방울토마토 수집 데이터는 초장, 엽수 등 생육 데이터와 온실 환경 데이터를 포함하여 총 38개의 분석 변수를 구성하였다(표 6, 표 7). 이후 변수와 착과수 간의 상관관계를 분석하여 주요 변수 후보를 탐색하였으며, 상관관계 분석 결과 재배 주차, 초장, 누적 외부일사량, 적산온도를 주요 변수를 선정하였다(그림 4, 표 8).

표 5. 오이 생산량 예측 모델 개발을 위한 변수 정의

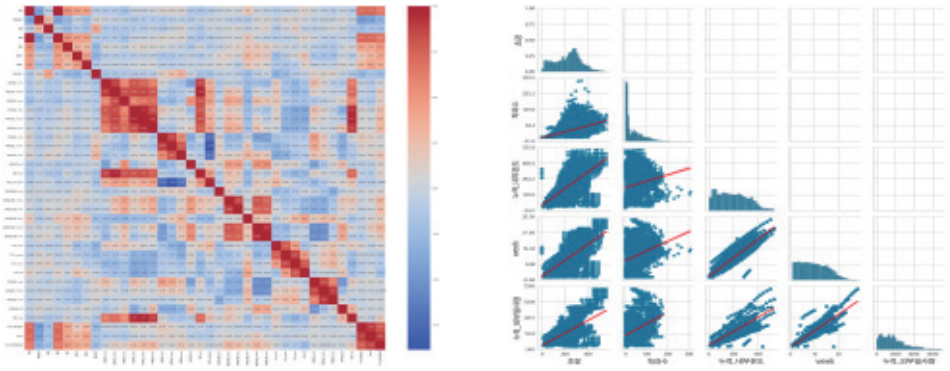
구 분	변수명	단위	데이터 타입	조사 기준
독립변수	정식 후 주차	number	integer	정식일로부터 경과한 주차
독립변수	내부온도	℃	float	내부온도의 1시간 단위 평균값
독립변수	내부상대습도	%	float	내부상대습도의 1시간 단위 평균값
독립변수	내부이산화탄소농도	ppm	float	내부이산화탄소농도의 1시간 단위 평균값
독립변수	외부일사량	W/m^2	float	외부일사량의 1시간 단위 평균값
종속변수	오이 수확량	ea	float	주차별 표본 당 수확량의 평균값

표 6. 방울토마토 수집 데이터 속성

분류	항목	단위	조사 방법
생육	초장	cm	지표면에서 생장점까지의 길이
	엽수	개	개화화방 아래 모든 엽의 수
	엽장	cm	개화화방 아래 2번째 엽의 길이
	엽폭	cm	개화화방 아래 2번째 엽의 잎폭 길이
	줄기굵기	mm	개화화방 아래 2번째 엽의 1~2cm 아래 가장 굵은 줄기 굵기
	화방높이	cm	생장점에서 개화화방이 갈라지는 지점까지의 거리
	착과수	개	전체 열매 개수
	수확수	개	수확한 열매수
환경	외부온도	℃	1시간 단위 수집(센서 계측)
	외부습도	%	1시간 단위 수집(센서 계측)
	외부일사량	W/m^2	1시간 단위 수집(센서 계측)
	내부온도	℃	1시간 단위 수집(센서 계측)
	내부습도	%	1시간 단위 수집(센서 계측)
	내부일사량	W/m^2	1시간 단위 수집(센서 계측)
	CO_2 농도	ppm	1시간 단위 수집(센서 계측)

표 7. 방울토마토 생산량 예측을 위한 주요 분석 변수 목록

구분	변수명	단위	조사 기준
생육	초장	cm	지표면에서 생장점까지의 길이
	엽수	개	개화화방 아래 잎의 수
	엽장	cm	개화화방 아래 두 번째 잎의 길이
	엽폭	cm	개화화방 아래 두 번째 잎의 폭
	줄기굵기	mm	개화화방 아래 1~2cm 지점 줄기 굵기
	화방높이	cm	생장점에서 개화화방까지의 거리
	착과수	개	개체당 전체 열매 개수
	수확수	개	개체당 수확된 열매 수
환경	내부온도_mean	℃	온실 내부 온도의 1시간 평균값
	내부온도_min	℃	온실 내부 온도의 1시간 최소값
	내부온도_max	℃	온실 내부 온도의 1시간 최대값
	내부습도_mean	%	온실 내부 습도의 1시간 평균값
	내부습도_min	%	온실 내부 습도의 1시간 최소값
	내부습도_max	%	온실 내부 습도의 1시간 최대값
	내부 CO_2_mean	ppm	온실 내부 CO_2 농도 평균값
	내부 CO_2_min	ppm	온실 내부 CO_2 농도 최소값
	내부 CO_2_max	ppm	온실 내부 CO_2 농도 최대값
	외부일사량_mean	W/m^2	외부 일사량의 평균값
	외부일사량_min	W/m^2	외부 일사량의 최소값
	외부일사량_max	W/m^2	외부 일사량의 최대값
파생	내부 VPD	kPa	(SVP×(1-내부습도)/100)/1000
	외부 SVP	kPa	610.78×exp((내부온도×17.2694)/(내부온도+233.3))
	내부온도_dif	℃	내부온도 최대값과 최소값의 차이
	외부온도_dif	℃	외부온도 최대값과 최소값의 차이
	CO_2_dif	ppm	CO_2 최대값과 최소값의 차이
	누적 외부일사량	W/m^2	주 평균 외부일사량의 누적값
	적산온도	℃	주 평균 내부온도의 누적값
	week	주	정식일로부터 경과한 주차



[상관관계 분석] [주요 변수 간 상관관계 분석]

그림 4. 방울토마토 변수간 상관관계 분석

표 8. 방울토마토 생산량 예측 모델 개발을 위한 변수 정의

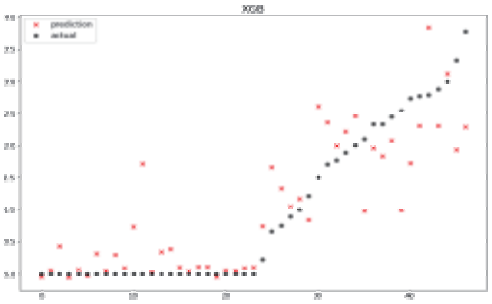
구 분	변수명	단위	데이터 타입	조사 기준
독립변수	정식 후 주차	number	integer	정식일로부터 경과한 주차
독립변수	초장	cm	float	지표면에서 생장점까지의 길이
독립변수	누적 외부일사량	W/m2	float	주 평균 외부일사량의 누적값
독립변수	적산온도	℃	float	주 평균 내부온도의 누적값
종속변수	착과수	ea	float	개체당 전체 열매 개수

입력 변수는 MinMaxScaler를 이용하여 정규화하였으며 결측치 제거 및 보정 등 데이터 전처리를 수행하였다. 생산량 예측 모델 개발을 위해 회귀 기반 모델과 앙상블 기반 모델을 적용하여 모델 성능을 비교하였다. 오이 수확량 예측 모델 분석 결과 XGBoost 알고리즘이 가장 우수한 예측 성능을 나타냈으며 GBM, LightGBM, Random Forest, Lasso, Ridge, ElasticNet, Linear의 순서로 예측 성능이 높은 것으로 나타났다. XGBoost 기반 모델의 성능 평가 결과 평균절대오차(MAE)는 0.448, 결정계수(R^2)는 0.714로 나타나 스마트팜 환경 데이터를 활용한 오이 수확량 예측 모델의 적용 가능성을 확인하였다. 특히 재배 주차, 이산화탄소 농도, 외부 일사량 등의 환경 변수가 수확량 변화와 상대적으로 높은 기여도를 나타내어 생산량 예측에 중요한 변수로 작용하는 것으로 확인되었다(표 9, 그림 5).

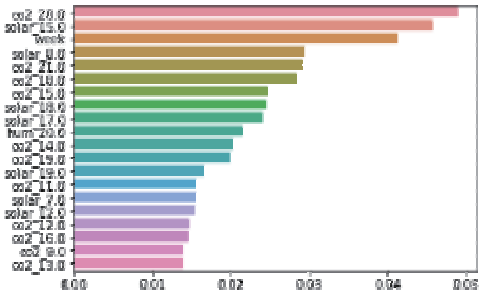
최근 연구에서도 머신러닝 기반 분석 모델이 작물 생산량 예측에서 높은 성능을 보이는 것으로 보고되고 있으며, Random Forest, Gradient Boosting 등 앙상블 기반 알고리즘이 농업 생산량 예측 연구에 효과적으로 활용되고 있는 것으로 보고된 바 있다(Odah 등, 2025; Mancer 등, 2024).

표 9. 오이 수확량 예측 모델 성능 비교

알고리즘	MAE	R^2
LinearRegression	1.24	-5.379
Ridge	0.719	0.407
Lasso	0.711	0.423
Elasticnet	0.731	0.359
Random Forest	0.455	0.670
GBM	0.446	0.697
XGBoost	0.448	0.714
LightGBM	0.47	0.694



[XGBoost 모델 예측]



[변수 중요도]

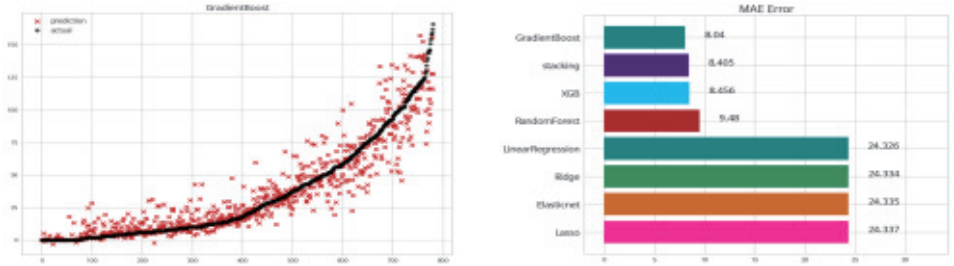
그림 5. XGBoost 모델 예측 성능 및 변수 중요도

방울토마토 생산량 예측 모델 분석 결과 Gradient Boosting 알고리즘이 가장 높은 예측 정확도를 나타냈다. Gradient Boosting 기반 모델의 성능 평가 결과 평균절대오차(MAE)는 8.04, 결정계수(R^2)는 0.88로 나타나 비교적 높은 예측 성능을 보였다(표 10, 그림 6). 분석 결과 초장, 누적 외부 일사량, 적산온도 등 생육 및 환경 변수는 착과수 변화와 밀접한 관계를 나타내었으며, 특히 누적 일사량과 적산온도는 생산량 예측에 중요한 환경 변수로 나타났다.

온실 환경 데이터를 활용한 작물 생산량 예측 연구에서도 누적 일사량, 온도 등 환경 변수와 작물 생육 지표가 생산량 예측에 중요한 변수로 작용하는 것으로 보고된 바 있다(Lin 등, 2019).

표 10. 방울토마토 생산량 예측 모델 성능 비교

알고리즘	MAE	R^2
LinearRegression	24.33	0.29
Ridge	24.33	0.29
Lasso	24.34	0.29
Elasticnet	24.33	0.29
Random Forest	9.48	0.86
GradientBoost	8.04	0.88
XGBoosting	8.46	0.87
Stacking	8.41	0.86



[XGBoost 모델 예측] [모델 별 성능평가]

그림 6. 방울토마토 생산량 예측 모델 성능 평가

<시험3> 생산량 예측 모델 시스템 구현

<시험 2>을 통해 개발된 생산량 예측 모델은 자체 운영 중인 정보시스템인 “경기도스마트팜데이터활용서비스(GSDUS)”에서 구현하였다. 환경 데이터는 스마트팜 센서 API를 통해 실시간으로 수집되도록 하였으며 수집된 데이터는 시스템에 자동으로 업데이트되도록 구성하였다. 또한 생육 데이터는 주 1회 엑셀 파일 형태로 업로드되며 환경 데이터와 재배 주차 기준으로 통합되어 생산량 예측 모델의 입력 변수로 활용되도록 하였다. 예측 모델 구동에 필요한 파생변수는 시스템 내부에서 자동 생성되도록 구현하였다. 주차별 평균 누적 외부 일사량과 주차별 평균 적산 온도 등 모델 입력 변수에 필요한 파생 변수를 생성하고 결측치 보정, 단위 변환 등 데이터 전처리 과정은 시스템 내부 로직을 통해 처리되도록 하였다.

예측 모델 연계 및 서비스 프로세스는 환경 데이터의 자동 수집과 생육 데이터 업로드를 기반으로 수행된다. 수집된 데이터는 전처리 과정을 거쳐 병합되고 모델 입력 변수로 변환된 후 생산량 예측 모델이 구동되어 재배 주차별 생산량 예측 결과가 생성된다. 예측 모델은 Python 기반 머신러닝 모듈로 구현하였으며 기존 정보시스템과의 연계를 위해 JAVA 기반 API를 통해 모델 호출 및 결과 반환이 이루어지도록 시스템 구조를 설계하였다(그림 7). 예측 모델 결과는 재배 기간 동안의 생산량 변화 추세를 주차별 그래프로 확인할 수 있도록 시각화하였다. 오이와 방울토마토 생산량

예측 결과는 재배 기간 중 생산량 변화 경향을 그래프 형태로 제공하여 생산량 변동 추이를 직관적으로 확인할 수 있도록 구현하였다(그림 8).

이러한 생산량 예측 결과의 시각화와 서비스 제공은 데이터 기반 재배 의사결정 지원 기술의 중요한 요소로 제시되고 있으며, 스마트농업 의사결정 지원 시스템 구축을 통해 재배 관리 및 생산 계획 수립에 활용될 수 있는 것으로 보고되고 있다(Gong 등, 2021).

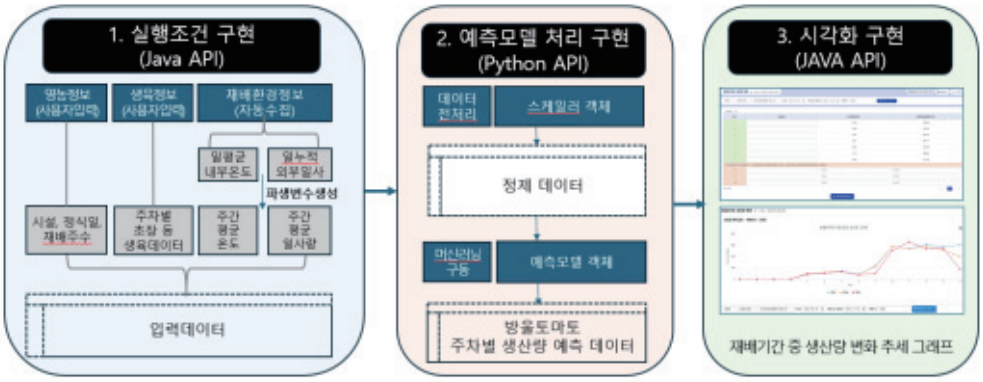


그림 7. 생산량 예측 서비스 구현 체계

[오이 생산량 예측 모델] [방울 토마토 생산량 예측 모델]

그림 8. GSDUS 시스템에 구축한 생산량 예측 시각화 모델

본 연구 과정에서 축적된 데이터 분석 기반을 활용하여 2025년에는 시설 가지 형상 특성 분석을 위한 이미지 기반 분석 연구를 추가로 수행하였다. 해당 연구는 데이터 기반 생육 분석 기술 고도화를 위해 후속 연구사업(지역특화작목 연구개발 지원 사업, ’26~’28)을 통해 지속적으로 추진할 예정이다.

4. 적 요

본 연구는 스마트팜 환경 및 생육 데이터를 활용하여 시설원예 작물의 생산량 예측 모델을 개발하고 이를 정보시스템과 연계한 서비스 형태로 구현하기 위해 수행되었다.

가. 데이터 기반 분석을 위해 데이터 표준화 및 데이터 관리 체계 구축을 수행하였으며, 단어, 용어, 도메인, 코드 사전으로 구성된 데이터 표준사전 구축 및 데이터 레이크, 데이터 웨어하우스, 데이터 마트 구조의 데이터 파이프라인을 설계하여 스마트팜 데이터 분석 환경을 구축하였다.

나. 재배 주차, 내부온도, 내부상대습도, 내부 이산화탄소 농도, 외부 일사량을 독립변수로 활용하여 XGBoost 기반 오이 수확량 예측 모델을 구축하였고, 그 결과 평균절대오차(MAE) 0.448, 결정계수(R^2) 0.714의 성능을 보였다. 방울토마토 예측 모델 개발을 위해 재배 주차, 초장, 누적 외부일사량, 적산온도를 활용하였고 Gradient Boosting 기반 모델에서 평균절대오차(MAE) 8.04, 결정계수(R^2) 0.88의 성능을 나타냈다.

다. 개발된 생산량 예측 모델은 자체 운영 중인 정보시스템인 '경기도스마트팜데이터활용서비스(GSDUS)'에 연계하여 환경 데이터 자동 수집, 생육 데이터 업로드, 데이터 전처리 및 파생 변수 생성 과정을 자동화하였다. 예측 결과는 재배 기간 동안의 생산량 변화 추세를 주차별 그래프로 시각화하여 제공하도록 구현하였다.

라. 본 연구는 스마트팜 데이터를 기반으로 작목별 생산량 예측 모델을 개발하고 이를 서비스 형태로 구현한 데이터 기반 스마트농업 의사결정 지원 기술의 활용 사례로 활용될 수 있을 것으로 판단된다.

5. 인용 문헌

김성란, 최경락, 유영글, 황연현, 김영광, 김영순. 2023. 스마트팜에서 빅데이터 분석을 활용한 경남 토마토 농가의 생산성 향상 모델 개발. JOKRE. 21(2):75-92.

김승재, 여현. 2022. 스마트팜 기술 동향 및 표준화 방안. JKICS. 47(11):1965-1973.

노시영, 원진호, 김현종, 최인찬, 곽강수. 2020. 스마트 농업 활성화를 위한 농업 빅데이터 플랫폼 구축 방안 연구. JKITS. 5(5):915-923.

노희선, 이윤숙. 2020. 토마토 스마트팜 생육데이터와 수확량의 연관성 분석. 융복합지식학회논문지. 8(3): 17-25.

한석호, 장훈석. 2023. 스마트팜 생육환경 데이터 획득 및 분석. JKIIECT. 16(3):130-137.

Gong L., Yu M., Jiang S., Cutsuridis V., and S. Pearson. 2021. Deep learning based prediction on greenhouse crop yield combined TCN and RNN. Sensors. 21(13): 4537.

Lin D., Wei R., and L. Xu. 2019. An Integrated yield prediction model for greenhouse tomato. Agronomy. 9(12):873

Mancer M., Terrissa L.S., and S. Ayad. 2024. Machine learning-based prediction of Tomato yield in greenhouse environment. ICEIS. 117-128.

Odah K.A., Houetohossou S.C.A., Houndhi V.R., and R.L.G. Kakai. 2025. Machine learning techniques for tomato yield prediction: A comprehensive analysis. Smart Agricultural Technology. 12: 101067

Solanki R. 2025. Convergence of data engineering and agriculture for sustainable farming systems. WJARR. 26(3):2292-2301.

6. 연구결과 활용제목

○ 경기도 스마트팜 데이터 활용 서비스의 오이 생산량 예측 서비스(영농활용, 2023년)
○ 경기도 스마트팜 데이터 활용 방울토마토 생산량 예측 모델(영농활용, 2024년)
○ 머신러닝 알고리즘을 이용한 오이 수확량 예측 모델(저작권, 2025년)
○ 경기도 스마트팜 데이터 활용 시설 방울토마토 생산량 예측 모델(저작권, 2025년)

7. 연구원 편성

세부과제	구분	소속	직급	성명	수행업무	참여년도			
						'22	'23	'24	'25
머신러닝 알고리즘을 활용한 과채류 수확량 예측 모델 개발	**책임자**	**원예연구과**	**농업연구사**	**김혜형**	**세부과제 총괄**	-	-	○	○
	공동	원예연구과	농업연구사	박남원	과제총괄	○	○	-	-
	연구자	〃	〃	정현경	자료조사	○	○	-	○
	〃	〃	〃	이슬기	생육조사	-	-	○	○
	〃	〃	〃	안주연	자료조사	-	-	-	○
	〃	〃	농업연구관	심상연	연구자문	-	-	-	○
	〃	〃	〃	이지영	연구자문	○	○	○	-
	〃	〃	〃	김진영	방향제시	-	-	-	○
	〃	〃	〃	이수연	방향제시	○	○	○	-