제11회 「2023 빅콘테스트」 결과보고서

* 해당란에 ☑ 표시

참가분야	☐ 생성형AI 분야 ☐ 정형데이터 분석 분야 ☑ 빅데이터플랫폼 활용 분야	☐ 데이터신기술 분야 ☐ 비정형데이터 분석 분야	
세부리그 *해당시 체크	☐ 어드밴스드 리그 *정형데이터 분석분야에 한함(선택)	☐ 스타터 리그	
	☑ 지정주제 리그 *빅데이터플랫폼 활용분야에 한함(선택)	☐ 자유주제 리그	
개인/팀여부	☐ 개인 ☑ 팀(총 4명)	**개인/팀명**	베리베리
지도교사명	–		
대표ID	alice010315@naver.com		

결과보고서 작성 안내 사항	
목차	

Ⅰ. 개요 및 문제 정의

1. 배경

스마트팜이란 '시간과 공간의 제약 없이' 자동으로 원격에서 생육환경을 관측하고 적절히 최적의 상태로 제어하는 과학 기반의 농업방식을 의미한다. 농촌 인구 고령화에 따른 인력 및 경지면적의 감소, 농가소득의 정체 등의 문제가 대두됨에 따라 국내 스마트팜 규모 및 수요가 증가하고 있다. 기존의 '경험 기반 농업'이 농부의 직관적인 경험 및 노하우에만 의존한 의사결정을 요구했다면, 스마트팜과 같은 데이터 기반 농업 시스템은 전산화된 시스템을 통해 부족한 인력 문제를 해결하고 농산물 생산성을 향상시키는 잠재력을 지니고 있다.

딸기는 재배 기간이 길고 요구되는 노동력이 많은 작물이며, 2020 년 기준으로 우리나라 전체 채소 생산액의 10.9%를 차지해 채소 작물 중 가장 재배 규모가 크다. 그러나 육묘, 수확, 선별에 들어가는 노동, 농업 인력의 노령화 및 감소 등의 이유로 재배 면적이 점차 줄어드는 추세이다. 이를 해소하기 위해 스마트팜 등의 선진 시스템을 적극적으로 도입할 필요가 있다.

2021 년을 기준으로 스마트팜 도입 시설원예 농가수는 858 호이다. 그러나 시설원예의 데이터 수집 및 분석 주체를 살펴보면 자가의 비중이 81.1% 로 매우 높게 나타나, 외부로의 데이터 이동, 데이터의 공유 및 분석이 잘 이루어지고 있지 않음을 알 수 있다. 그 원인으로 연구자들은 첫째, 농업인들은 데이터 사용 목적이 불분명하고 데이터로 수행되는 작업이 충분히 이해되지 않으면 데이터를 공유하지 않으려는 경향이 강하며 (Borrero & Mariscal, 2022) 둘째, 생육, 환경, 유통, 소비 등 각종 농업 데이터를 공공기관·기업·다른 농가 등 외부에 제공해야 하는 이유와 제공 시의 혜택을 잘 알고 있지 못한다 (변재연, 2022)는 점을 꼽았다. 이러한 현상은 농업 데이터 활용 활성화를 저하시키는 요인이 되고 있다.

물론 제도적 측면에서 데이터 사용 권한에 대한 규정 및 데이터 제공에 따른 합당한 보상을 지불하는 부가 시스템을 마련하는 것도 중요하다. 그러나 근본적으로는 **농업인들이 농업데이터를 지속적으로 제작하여 외부에 제공하도록 납득할 수 있는 합리적인 근거를 제공할 수 있어야만 한다**. 따라서 본 팀은 "고도화된 스마트팜 시스템이 필요하다"는 문제 의식 하에, 농가의 수입과 직접적인 연관이 있는 착과수 예측 모델을 개발하고 이를 발전시킬 다양한 아이디어를 제안함으로써 목표 달성에 기여하고자 한다.

2. 선행 연구 검토

최근 들어 **착과수를 비롯한 딸기의 생육 변수를 예측하는 모델을 구축하고자 하는 국내 연구가 활발**히 이루어지고 있다. 백진동 (2020)은 온실의 내/외부 환경 정보를 이용하여, 현재의 상태가 딸기 생육에 적합한 환경인지의 여부를 예측하는 모델을 구축하였다. 김나은 외 (2022)는 온실에서 재배되는 수확량. 온도, 습도, CO2 와 같은 환경적인 매개변수와 생리학적인 '설향'의 잎의 길이, 꽃과 열매의 수, 엽록소 함량 등의 매개변수(wide) 를 함께 사용하여 딸기 생산량을 예측하였다. 이인하 외 (2021)은 분석을 통해

딸기 착과수와 초장, 엽수, 액아수와 같은 생육 변수 간의 관계를 파악하였고, 김은완 외 (2022)는 딸기 이미지 데이터를 이용하여 딸기 생육 작기를 시각적으로 구분하는 모델을 설계하였다. 이서희 외 (2022)는 상관분석을 통해 착과수와 관련이 있는 생육 변수 및 환경 변수를 규명하였다.

딸기 외의 유사 농작물 토마토에 대해서도, 노희선 & 이윤숙 (2020)은 초장, 생장길이, 엽수, 엽길이 등의 생육 변수를 이용하여 토마토 수확량을 예측하는 모델을 구축하였으며, 홍성은 외(2020)는 생육 정보와 내부 환경 정보를 함께 사용하여 생산량 및 성장량 예측 모델을 구현하였다. 이세연 외 (2023)의 연구에서는 양방향 LSTM 을 이용하여 토마토 생산량을 예측하였는데, 본 팀은 해당 연구를 참고하여 (하반기 20%에 해당하는 착과수를 예측해야하는) 단방향 LSTM 을 설계하였다.

다만, 이와 같은 선행 연구들을 검토하면서 몇 가지 아쉬움이 남을 수밖에 없었다. 첫째, 모델의 측면에서, **분석에 사용되는 모델의 다양성이 다소 떨어지는 모습**이 관찰되었다. 대부분의 논문은 기본적 회귀 모델로 Ridge Regression 을 사용하였으며, 통계적 해석이 아닌 데이터 기반 모델을 구축한 몇 안되는 경우도 LSTM 만을 이용하여 다른 최신 머신러닝 모델 등과의 성능 비교가 이루어지지 않았다. 나아가 모델의 예측 결과에 대한 깊이있는 해석이 부족했다고 판단되었다.

둘째, **생육 변수와 환경 변수의 구분 없는 사용, 또는 한쪽에만 편중된 해석**이 완연했다. 생육 변수와 환경 변수는 명백히 입장이 다르다. 생육 변수는 착과수와 마찬가지로 **환경 변수의 영향을 함께 받는 또 다른 예측 대상**이다. **즉 (내부 환경 변수와는 달리) 농업인이 자율적으로 조절할 수 있는 변수가 아닌, 복잡한 환경적 영향을 받는 변수**이기 때문에, 생육 변수들을 사용한 예측 모델은 '*감독 모델*', 환경 변수들을 사용한 예측 모델은 '*최적 환경 모델*'에 가깝다. 이렇게 **특성이 다른 두 개의 변수를 하나의 모형 구축에 사용하는 것은 예측 성능 자체를 높일 수 있을지는 몰라도, 농업인들에게 명확한 (단일 목적의) 인사이트를 제공해주기는 어렵다**. 반대로 이 둘은 방식은 다르지만 둘 다 착과수와 분명한 관련이 있기 때문에 동등한 정도로 이 관계를 해석하는 것이 필요하다고 본다.

셋째, **예측 모형의 의사 결정에 '농부의 감'이 전혀 반영되지 않았다**. 물론 모델 구축에 있어서 주어진 변수 외의 외부 변수를 제작하는 것은 매우 조심스럽게 이루어져야 하는 과정이다. 그러나 농업은 특히나 사람만이 갖고 있는 인사이트가 중요한 분야라고 생각한다. 데이터를 분석하는 주체와 달리 농업인들은 축적된 경험을 가지고 있고, 이전의 수확량 및 수확 패턴에 대한 기억을 바탕으로 착과수를 좀 더 유동적으로 예측할 수 있다. 만약 (도메인 지식, 또는 분석 외 데이터를 기반으로 한) **'충분한 근거가 있는' 변수를 도입하는 것이 가능하다면, 이와 같은 인간의 경험적 지식을 반영할 수 있는 변수를 제작하는 것도 합당**해보인다.

마지막으로, **추가 연구에 대한 아이디어 제시가 부족**했다. 대부분의 연구들은 주어진 데이터를 통해 착과수를 단편적으로 예측하는 데에 주 의의를 두었다. 그러나 이와 같은 모델들은 (본 팀의 모델 역시 갖는 한계점이기도 하지만) 최적 데이터 또는 학습에 사용된 데이터가 아닌, **스마트팜 농사에 처음 진입한 '초보자'들의 농작물 데이터를 입력 받았을 때 제대로 대처하지 못할 가능성**이 높다. 또한 연구들에 언급된 추후 연구 방향성은 주로 생육 변수 값을 좀 더 정확하게 예측하고, 최적 환경과 현재 환경을 비교하여 내부 환경 조성에 대한 생육 가이드라인을 제시하는 것을 목표로 하고 있었다. 그러나 생육 예측 모델 개발에 있어서 현 패러다임을 유지하는 것은 *생육 변수 및 환경 변수 간의 복잡하고 통합적인 관계를 규명* 하거나, 단순 상관관계 기반이 아닌 *인과관계 기반의 (현재 내부 온도를 이와 같이*

수정한다면, 미래의 착과수가 이렇게 변화할 것이다) 의사결정을 돕는 모델을 개발하는 것 등에는 큰 도움이 되지 못할 것으로 사료된다. 즉 **보다 상위의 목표 달성을 위한 연구 아이디어가 부족**한 것이 실정이었다.

3. 목적 및 필요성

본 팀이 구축한 착과수 예측 모형 또한 기본적으로는 기존 연구와 동일하게 단편적인 데이터를 입력 받아, 생육과 관련된 정보를 이용하여 착과수를 잘 예측하는 것을 목표로 한다. 다만 선행 연구에서 느낀 여러가지 아쉬움을 보완하기 위해, 본 팀은 우선 **모델 구축에 있어서 다양한 모델(통계적 회귀 기반, Auto ML 기반, LSTM 기반) 적합을 시도하고 이들의 성능 및 예측 결과를 비교**하였다. 전역적 또는 지역적 단위의 통계 모델 적합은 블랙박스 모델에서는 부족할 수 밖에 없는 설명력 확보를 제한적으로나마 가능하게 하며, ML 및 DL 모델은 예측 성능을 높여주는 역할을 충실하게 수행하였다. 그리고 **현재의 모델로는 수행하기 어렵지만 연구될 필요성이 있는 연구 문제들을 자기 회귀, 구조방정식, 제약 조건이 있는 최적화, 인과 추론과 같은 각종 통계학 분야와 결합시켜, 최대한 다양한 아이디어를 보고서 전반에 제시**하고자 노력하였다.

나아가 변수를 선택하는 과정에서 **생육 변수와 환경 변수를 구분**하여, 착과수 - 생육 변수 (사후 분석), 착과수 - 환경 변수 (모델링) 각각에 개별적으로 접근하였다. 특히 환경 변수의 경우 외부 환경 변수에 일정한 제어 변수 처치를 하여 결과적으로 얻어진 것이 내부 환경 변수라고 판단하였고, 따라서 **농업인이 직접 제어가 가능한 내부 환경 변수만을 이용하여 최적 환경 하의 착과수를 예측하는 모델을 적합**하되, *각 내부 환경 변수가 외부 환경 변수와 어떤 관련이 있는지를 추가적인 사후 분석을 통해 살펴보았다*. 본 팀은 또한 **도메인적 지식 및 기타 데이터 분석을 통해 얻을 수 있는 인사이트를 활용**하여 GDD, 시기 변수와 같은 새로운 변수를 추가함으로써 모델의 예측 성능을 향상시켰다.

본 프로젝트의 궁극적 목적은 다음과 같다. 첫째, 성능이 우수하며 사전 지식과 상호작용 할 수 있는 착과수 예측 모형을 구축함으로써 **농업인들에게 정보를 제공하고, 그들이 양질의 농업 데이터를 제공할 필요성을 느낄 수 있도록 돕는다**. 둘째, **스마트팜 데이터 플랫폼 관리자들에게 오픈된 관련 데이터 축적의 중요성을 강조**한다. 셋째, **연구자들에게 다양한 후속 연구 아이디어를 제시**함으로써 해당 산업의 발전에 기여한다. 마지막으로, 데이터 산업에 종사할 미래 인력으로서, **스마트팜 데이터 분석을 통한 의사결정 분야에 갖는 (스스로의) 관심을 증진**시킨다.

Ⅱ. 문제 수행 준비

1. 분석 데이터 선정

시기별 착과수 예측 모델 개발이 목적인 만큼 모델 구축에 사용될 주요 데이터는 네이버 스마트팜 빅데이터 플랫폼에서 제공하는 '**딸기 착과수 최적환경 데이터**'이다. 또한 본 팀은 **딸기 생육 정보와 관련된 25 개의 csv 데이터를 활용**하여, 사후분석을 통해 *딸기와 관련된 다른 생육 정보와 착과수 간의 관계를 유추하고 딸기 생육을 위한 전반적 인사이트를 제공* 하고자 한다.

네이버 스마트팜 빅데이터 플랫폼에서 제공하는 데이터 형태는 크게 csv 와 이미지의 2 가지로 분류된다. 본래 모델링 단계에서 본 팀은 **플랫폼에서 제공하는 이미지 데이터 셋과 외부 데이터를 활용**하여 *해당 농가의 시기별 딸기 생육 단계를 예측하는 1 차 모델을 구축* 하고, 이를 바탕으로 *생육 단계별 착과수 예측 2 차 모델을 적합* 하고자 하였다. 다만 csv 데이터와 이미지 데이터의 데이터 수집 시기가 상이함(csv 는 1 월부터 6 월, 이미지는 1 월과 12 월)을 고려하여, 본 모델 구축에 있어서는 **csv 형식의 생육 데이터만 사용**하기로 결정하였다.

나아가 외부 데이터의 경우 기존에 사용하기로 하였던 **AI HUB 데이터**의 용량(약 40GB)으로 인해 환경적 제약이 발생하여 사용하지 않기로 결정하였다. AI HUB 데이터에서 보여지는 생육 작기 정보를 통해 본 데이터의 생육 작기 정보를 추론하는 것이 목적이었기 때문에, **본 데이터에 EDA 및 클러스터링 기법을 사용하여 미사용 데이터를 보강하고 작기와 관련된 정보를 추론하는 과정으로 이를 보완**하였다. 이처럼 착과수 예측 모델에 사용될 초기 아이디어 및 데이터의 구성 및 활용에 대한 구체적인 설명은 추후 섹션에서 논의하도록 하겠다.

2. 변수 정의 및 선택

가. 변수 정의

분석에 주로 사용될 딸기 착과수 최적환경 데이터는 66 번 Zone 에서 측정된 시계열 자료로, header 를 제외한 26063 개의 time point 에 대해 착과수를 포함한 101 개의 변수값이 기록되어 있다. 사후 분석에 사용될 25 개의 데이터셋은 착과수 대신 각기 다른 생육 정보를 포함하고 있다. 이처럼 딸기 자체에 대한 정보를 담고 있으면서 각 csv 에 대해 상이한 변수를 '**생육 변수**'라고 명명하였다. 착과수, 당도, 엽수 등이 이에 해당한다.

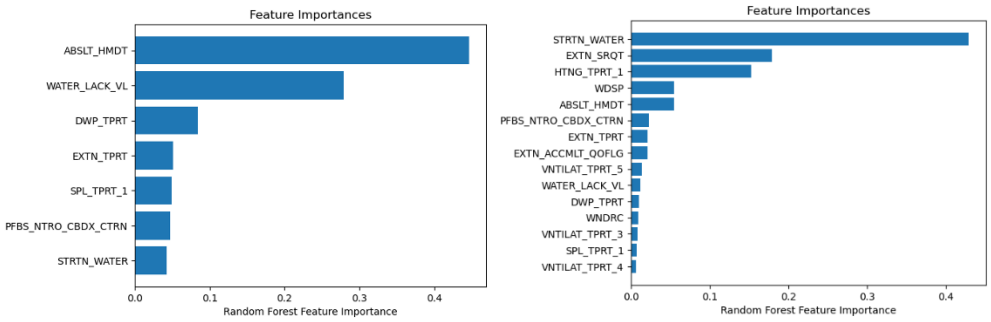
생육 변수와 시간, Zone 번호를 제외한 나머지 100 여개의 '**환경 변수**'는 크게 3 가지의 하위 카테고리로 분류될 수 있다. 첫째, 외부 온도나 습도, 풍향과 같이 농부가 조절하지 못하는 환경적 '**외부 변수**'들이다. 둘째, 자연 변수를 제외한 나머지 연속형 변수들이 속하는 카테고리로, 농부가 특정 처치를 통해 인공적으로 구축한 환경을 나타내는 '**내부 변수**'가 있겠다. 결과 변수는 (1) 내부 습도 및 온도와 같이

자연 변수 값 (외부 습도, 온도)을 설비 (분무장치, 난방기)를 통해 변화시킨 결과물 과 (2) 천창 개도율, 난방 온도와 같이 *농부가 특정 환경을 조성하기 위해 의도적으로 설정한 결과물* 로 세분화된다. 마지막으로 온실 내 환경을 관리하는 설비들의 작동 모드 및 여부를 나타내는 범주형 변수인 '**제어 변수**'들이 존재한다.

본 팀의 모델이 지닌 차별성은 설명 변수들을 구분 없이 무분별하게 사용하지 않고, 해석 단계를 고려하여 **내부 변수**만을 가지고 학습되었다는 점이다. 농부가 직접적으로 조절할 수 있는 수치인 내부 변수만을 가지고 생육 환경과 착과수 간의 관계를 설명한다면 보다 현실적으로 도움이 되는 모델 적합이 가능할 것이라고 판단하였다. 나아가 주요 내부 변수가 '*어떤 외부 변수에 어떤 제어 변수 처치에 의해 도출된 값인지*'를 추가적으로 분석하여 **data-based 생육 가이드라인에 필요한 토대**를 만들고자 하였다.

나. 변수 선택

우선 Raw data 는 설명변수의 수가 지나치게 많아 모델 적합에 바로 사용되기는 부적절한 구조를 띄고 있다. 이에 해당 조는 **여러 기준을 고려하여 분석에 사용될 총 7 개의 설명 변수를 구성**하였다. 먼저 모든 time point 에 대해 동일한 값을 공유하는, 또는 1~2 개의 값만 (데이터 입력 시 발생한 오류라고 사유된다) 다르게 갖는 열을 삭제하여 61 개의 열만을 남겼고, 이중 사후 분석에만 사용될 범주형 변수를 제거하자 32 개의 열이 남았다. 이후 (1) 착과수와의 **상관관계**가 낮게 나타나는 변수 제거 및 보류 (2) **Lasso regression** 을 통한 feature selection (3) **Random Forest** 를 통한 비선형 주요 변수 파악을 순차적으로 진행하여 착과수와 시간을 포함한 15 개의 변수를 선택하였다.

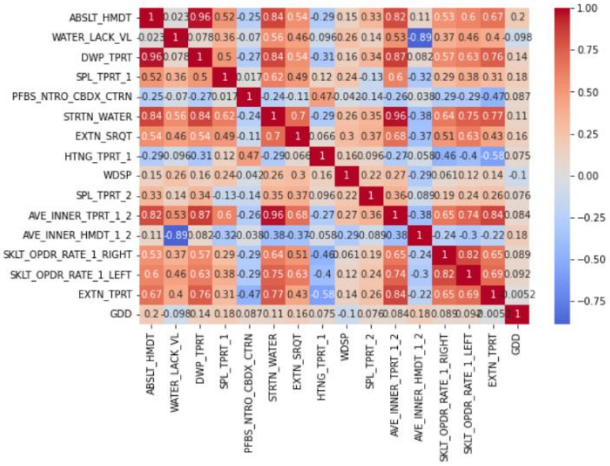


▲ 왼쪽은 상관관계가 낮은 변수를 제거 후 Random Forest 를 통해 구한 Feature Importance 로 전부 분석에 사용하였으며, 동일 과정을 오른쪽의 변수 제거가 없는 상태에도 적용한 후 일부만 분석에 사용하였다.

본 팀은 또한 사전 자료 조사를 통해 *딸기가 생육되는 장소의 내부 온도*가 생육 결과에 유의미한 영향을 미칠 것이라고 판단하였으나, feature selection 을 통해 선택된 변수 중 내부 온도를 직접적으로 나타내는 변수가 존재하지 않아 이를 간접적으로 대변할 수 있는 지표로서 **GDD**(Growing Degree Day)를 도입하였다. 작물의 성장을 추적하고 예측하는 데 도움이 되는 지표인 GDD 는 온도에 기반하여 계산되며, 특정 작물이 발달하는 데 필요한 열량 누적을 나타낸다. 온실 내부 온도 변수인 AVE_INNER_TPRT_1_2 와 딸기의 base temperature = 7 도(Mendonça et al., 2012)를 통해 아래의 수식으로 계산할 수 있다.

$$T_{\min} < T_{\text{base}} \text{ to } T_{\min} = T_{\text{base}}$$

$$GDD = \frac{T_{\max} + T_{\min}}{2} - T_{\text{base}}.$$

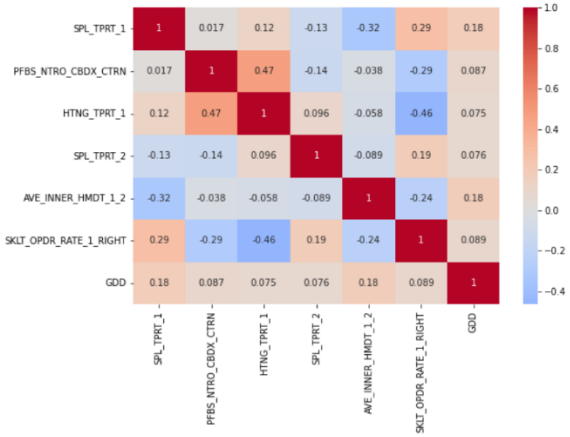


VIF

	feature	VIF
0	ABSLT_HMDT	18.541720
1	SPL_TPRT_1	23.205095
2	PFBS_NTRO_CBDX_CTRN	31.545807
3	EXTN_SRQT	4.141061
4	HTNG_TPRT_1	7.317536
5	WDSP	2.626620
6	SPL_TPRT_2	32.358585
7	AVE_INNER_HMDT_1_2	33.389030
8	SKLT_OPDR_RATE_1_RIGHT	5.517328
9	SKLT_OPDR_RATE_1_LEFT	6.947628
10	EXTN_TPRT	9.373780
11	GDD	1.111470

▲ 선택한 변수 전체의 상관관계 시각화 결과 (좌), 다중공선성 확인 결과 (우). 결과적으로 STRTN_WATER, AVE_INNER_TPRT_1_2, WATER_LACK_VL, DWP_TPRT 가 각기 다른 이유에 의해 분석에서 제외되었다.

선택된 변수들에 대해 상관분석 및 VIF 를 통한 다중공선성 확인을 진행하였다. 특정 변수와 *지나치게 높은 상관관계를 보인 변수*, 그리고 *분석 목적에 맞지 않는 외부 변수*를 제거하자 최종적으로 7 개의 설명 변수가 남았다. Feature selection 과정에서 중요하다고 파악된 자연 변수(절대습도, 외부온도, 외부일사량)들은 사후 분석에서 **결과 변수를 설명하는데 핵심적으로 사용될 예정**이다. 이렇게 구성된 변수들은 서로 양호한 상관 및 VIF 계수 값을 보이고 있다.

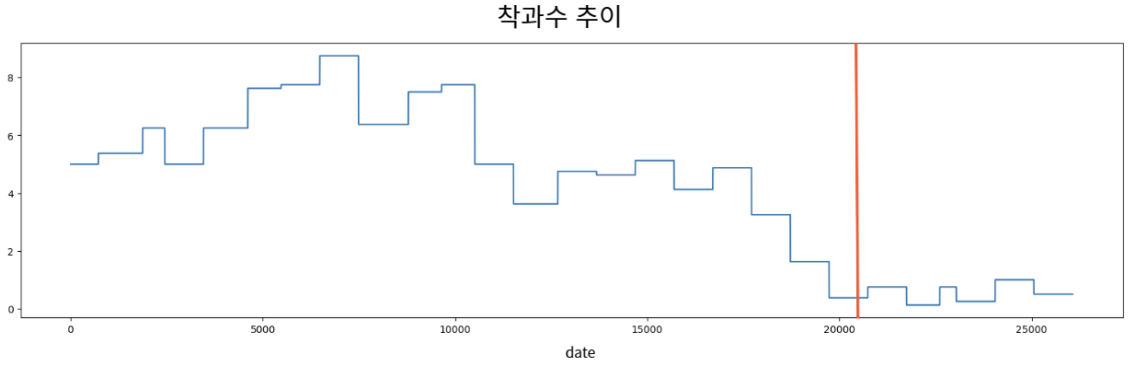


▲ 최종적으로 사용될 변수는 관측 지점 실내 이산화탄소 농도, 평균 내부 습도, 생육 온도 일수, 난방온도, 천창개도율, 공급온도 1, 공급온도 2 이다.

다음 단계에서는 회귀 및 Auto ML 모형에 '시간'의 개념을 도입하기 위해 생육 단계를 변수화한 과정을 설명한다. 착과수는 이 생육 단계 변수를 포함한 총 9~10 개의 설명 변수로 적합되었다.

다. 예측 대상

하반기 20%에 해당하는 착과수를 보다 정확하게 예측하는 것이 모델링의 주요 목표이다. 이때 아래의 계단형 그림을 보면 알 수 있듯이, 해당 데이터는 10 분 단위로 기록되어 있으나 동일 날짜에 대한 착과수는 전부 같은 값을 갖는다.



해당 데이터의 예측력은 **RMSE** 로 평가된다. RMSE 는 예측된 값과 실제 값의 차이를 통해 오차를 계산하는 것으로, RMSE 가 낮을수록 예측 성능이 더 좋다고 평가될 수 있다.

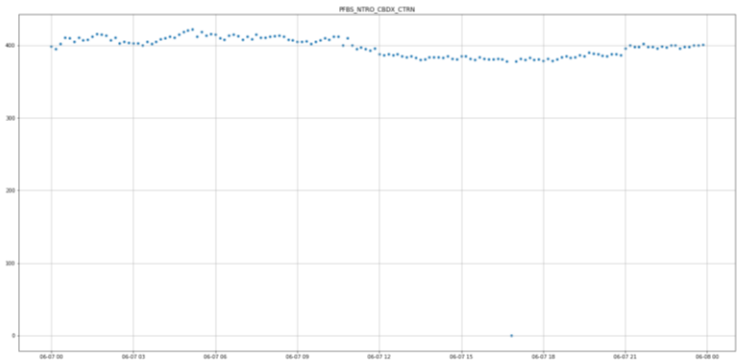
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

모델의 성능을 소폭 개선하는 방법의 일환으로, 본 팀은 데이터 형태의 특수성을 고려하여 **하루 동안 예측된 착과수의 일별 평균을 최종 예측값으로 반환**하는 과정을 거쳤다. *이로 인한 정보량의 손실은 사후 분석 및 예측 결과 해석을 통해 보완될 예정* 이다. 본 프로젝트에서는 시행되지 않았지만, 만약 이와 같은 변환 방식이 유의미한 성능 향상 효과를 보인다면 최종 예측값을 다른 대푯값 (중앙값, 최소값, 최대값, Q3 등) 으로 대체하며 성능을 비교해볼 수도 있을 것이다. 나아가 결과 해석에 있어서, *하루의 144 개 데이터를 각각 설명 변수로 두고 이들을 이용하여 최종 예측값을 설명하는 접근 방식* 역시 가능할 것으로 보인다.

3. 데이터 전처리

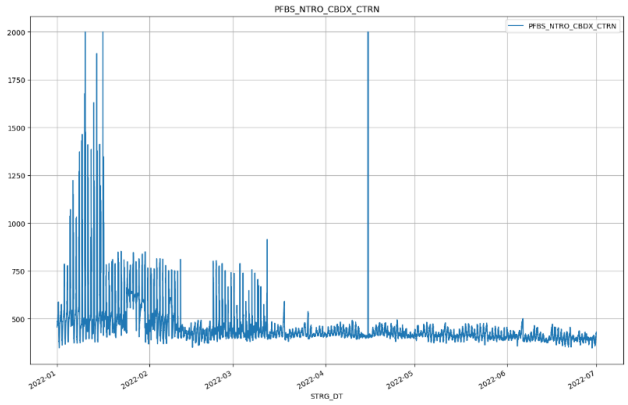
최종 선택된 총 6 개 (GDD 는 후에 생성된 변수로서 전처리가 필요치 않음)의 환경 변수에 대해 결측치 처리와 이상치 처리 그리고 정규화를 진행하였다. 데이터에 존재하는 결측치를 올바르게 처리하기 위해서는 각 설명 변수의 global 한 변화 양상 뿐만 아니라 **결측치가 존재하는 일자, 특히 결측치 부근의 time point 에 대한 변수의 local 한 변화 양상을 파악하는 것**이 중요하다고 판단하였다. 따라서 결측치가 존재하는 날의 데이터만을 따로 추출하여 변수별로 시각화한 후 각 변수의 특성을 살려 결측치를 채우고자 하였다.

착과수 데이터셋에서는 총 1 개의 time point (6 월 7 일 16 시 50 분)에 대한 결측치만이 확인되었으며, 분석에 사용할 각 변수의 6 월 7 일 데이터를 시각화한 결과 모든 변수의 결측치 부근에서 연속적으로 완만하게 변화하는 양상이 관찰되었다. 이에 *일반적인 보간법을 사용하여 앞뒤 데이터의 평균으로 결측치를 채워넣은 후* 나머지 데이터와 자릿수를 통일하기 위해 반올림 처리만을 거쳐주었다.



▲ 위의 사진은 6 월 7 일 16 시 50 분의 데이터를 0 으로 채운 후 시각화한 예시로, 전후 변화 양상을 고려하였을 때 보간법을 사용하는 것이 합당하다고 판단하였다.

데이터셋의 **이상치**는 (1) 직접 처리와 (2) 정규화(를 통한 간접 처리) 두 가지 방식을 사용하여 정제될 수 있다. 먼저 선택한 n 개 열에 대해 시간을 x 축, 설명변수를 y 축으로 한 그래프를 그려본 결과 대부분의 열에서 일정한 추세가 관찰되었다. 다만 예외적으로 PFBS_NTRO_CBDX_CTR 의 경우 한 단일 point 에서만 값이 크게 증가하는 이상치가 관측되었다. 해당 이상치는 앞 행과 다음 행의 평균으로 대체하여 더 정확한 예측을 할 수 있도록 보완해주었다.



▲ 4 월 중순의 단일 time stamp 에서 주변 점들과 크게 다른 값이 관찰되어, 이를 이상치로 판단하고 처리하였다.

마지막으로 대회 요강을 고려하여 하반기 20%를 test data 로 사용하기 위해 **randomize 없이 8:2 로 train data 와 test data 를 분리**하였고, 이후 각 변수들이 전부 동일한 비중(scale)으로 모델에 반영될 수 있도록 **정규화** 과정을 거쳐주었다. EDA 결과 많은 양의 이상치가 관찰되지는 않았으나, 문헌 조사 및 EDA 만으로 파악되지 않은 변수 자체의 특성이 존재할 것으로 판단하여 StandardScale 를 통한 정규화를 진행하였다. StandardScaler 는 각 변수에 대한 feature 값의 평균을 0, 표준편차를 1 로 간주하여 정규화한다. 파이썬 scikit-learn 라이브러리의 StandardScaler 를 사용하였다.

데이터 정리 후, 문제 특성 상 train data(이전 시기)로 적합된 모델이 test data(이후 시기)를 잘 예측하기 위해서는 '**시간**'이라는 특수 변수에 대한 고려가 필요할 것이라고 판단되었다. 더불어 착과수 데이터셋을 포함한 모든 사용 데이터들은 **동일 날짜에 대해 전부 동일한 값을 갖는 특이 구조**를 띄고 있다. 이를 위해 *모델링 각각에서 어떠한 모델 구축 및 분석 전략이 사용되었는지*가 뒤따르는 섹션에서 자세하게 논의될 예정이다.

Ⅲ. 문제 수행 절차

1. 생육 단계 구분선 결정

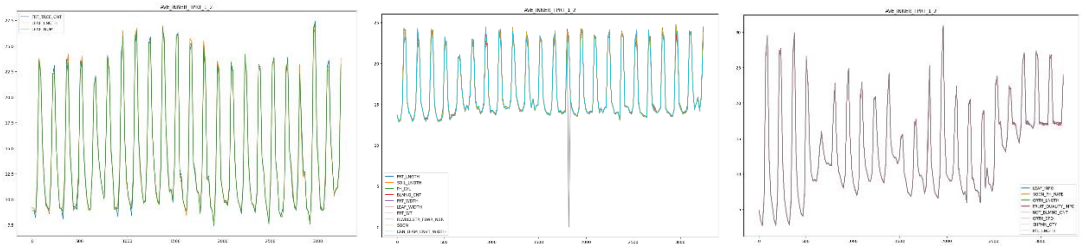
가. 변수 도입 및 EDA

분석에 있어서 가장 유의해야 하는 점 중 하나는 Test data 로 사용할 데이터가 하반기 20%의 생육 변수이며, 해당 단계에서의 변화 양상이 나머지 80%의 그것과는 상이한 구조를 띄고 있다는 사실이다. 본 팀이 가장 집중적으로 고민한 것은 '어떻게 하면 이러한 하반기 20%의 데이터를 잘 예측할 수 있을지'였고, 결론적으로 첫째, *overfitting 을 줄일 수 있는 모델을 사용* 하고 둘째, (LSTM 을 제외한 모델에서) **시간 또는 생육 시기를 변수화한 새로운 설명변수를 도입**하기로 결정하였다.

딸기는 기본적으로 정식기, 출뢰기, 개화기, 과실비대기, 수확기의 생육 단계를 거치며, 생육 시기 별로 조성해야할 환경이 조금씩 다른 것으로 알려져 있다. 나아가 동일한 수확기 내에 수확된 작물이라도, 과일은 기본적으로 수확기의 초, 중, 후반 중 언제 수확되었는지에 따라 품질 및 수확량이 미묘하게 다르다. 물론 난방 온도와 같은 변수를 통해 간접적으로 시간의 변화를 반영할 수 있겠지만, 이와 같은 '시간적' 사전 지식을 더 적극적으로 활용하기 위한 해결책이 '**시기 변수**'이다.

이 변수 제작을 위해 다양한 방식이 논의되었는데, 본 팀은 변수 도입의 타당성을 확보하기 위해 (test 의 편향이 생기는 것을 방지하기 위해) **예측의 대상이 되는 착과수가 아닌 다른 생육 변수의 변화 양상을 이용하여 변수를 제작** 해야 할 필요성을 느꼈다. 그러나 이는 *착과수와 변수 제작에 사용될 생육 변수가 같은 시공간적 환경에서 측정*되었음이 보장되어야 유의미하다. 즉 다른 생육 변수 데이터가 착과수 데이터와 동일 날짜에 동일한 단계를 공유하고 있는지를 먼저 파악해야 했다. 따라서 본 팀은 EDA 단계에서 20 여개의 데이터를 시각화 하였다.

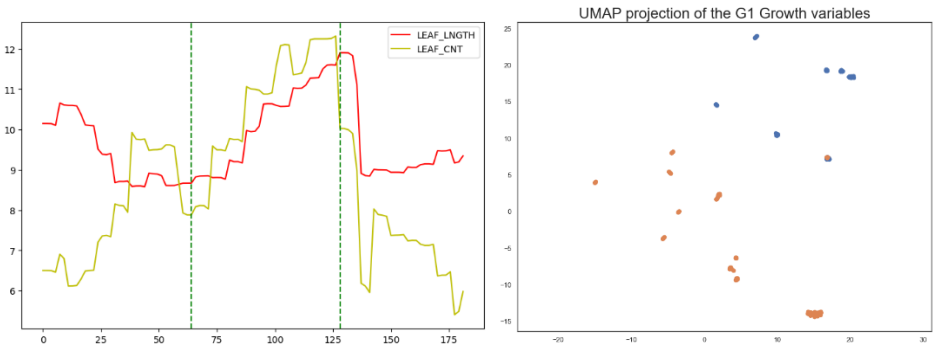
각 데이터의 *외부 온도, 습도, 풍속을 통해 동일 농가에서 측정된 데이터인지를 확인* 하였고, *평균 내부 온도, 평균 내부 습도 정보를 통해 동일한 생육 단계를 공유하고 있는지 확인* 하였다. 만약 이 정보들이 착과수 데이터와 동일하게 기록되어 있다면, 해당 데이터를 **착과수 데이터셋과 동일 작기를 공유하고 있다**고 분류하였다. 시각적 명확함을 위해 6 시간 단위로 데이터를 추출하여 Plotting 을 진행하였다.



결과적으로 24개의 생육 변수 데이터는 *총 3개의 그룹으로 분류되었다*. 위의 그림은 각 그룹에 해당하는 생육 변수 데이터의 평균 내부 온도 값을 그래프로 겹쳐 그린 것으로, 0으로 처리된 결측치를 제외하고는 값이 온전히 겹쳐지는 것을 관찰할 수 있다. 착과수 데이터와 동일 작기를 공유하고 있는 것으로 추정되는 데이터는 **LEAF_LNGTH와 LEAF_CNT**이다.

나. 시기 변수 제작

시기 변수 제작에는 총 3가지 방법이 사용되었다. 첫째, LEAF_LNGTH와 LEAF_CNT의 시간에 따른 값 변화를 시각화하여, **상관관계, 변곡점, 추세 변화를 고려한 '눈을 믿는' 구분 기준**을 결정한다. 알다시피 때로는 직관이 강력한 힘을 발휘한다. 둘째, LEAF_LNGTH와 LEAF_CNT에 대한 정보를 **DBscan, Mean-shift 알고리즘을 통해 군집화한다**. 군집이 명확하게 묶이도록 해주는 구분 기준이 있다면 이를 따른다. 마지막으로, **Graphical한 차원 축소 method인 UMAP을 이용하여 구분선을 탐색**한다. UMAP은 차원축소 기법이기 때문에, 3차원 이상이어야 2차원 평면 상에 시각화 하는 것이 의미를 가질 것으로 판단하여 단계 결정에 (착과수를 포함한) G1 그룹 전체를 사용하였다.



▲ EDA를 통한 시각화 결과 (좌)와 UMAP을 통한 시각화 결과 (우), 후자는 날짜 소그룹의 색을 변경해가면서 언제를 기준으로 다른 색깔 인덱스를 주어야 두 개의 색이 명확하게 분리되는 지를 확인하였다

결과적으로 **EDA를 통한 기준선1과 UMAP을 통한 기준선2의 2가지 후보가 제시**되었다 (군집화 알고리즘은 연속적인 날짜들이 하나의 군집에 묶이도록 분리해주지 못하였다). 기준선1은 총 181일의 데이터를 1일부터 64일, 65일부터 128일, 129일부터 181일의 3단계로 나누어준다. 기준선2는 1일부터 25일과 140일부터 181일을 하나의 단계로 묶고, 나머지 26일부터 139일을 동일 단계로 분류한다. 이 단계를 **cluster_n이라는 이름의 범주형 변수로 도입**하고, 모델 적합을 위해 one-hot encoding을 해주는 과정을 거쳤다.

2. 통계적 회귀 모형

가. 사용 모델

전통적 회귀 모형은 다른 블랙박스 모형에 비해 예측 성능이 떨어지나 강력한 설명력을 확보할 수 있다는 강점이 있다. 본 팀은 M*ultiple Linear Regression, Ridge Regression, Support Vector Regression* 과 같은 다양한 회귀 모델을 적합하였으며, 이들 중 가장 준수한 예측 성능 및 설명력을 지니고 있다고 판단되는 **GAM** 을 해당 섹션의 최종 모델로 선정하였다.

GAM 은 기존의 다중선형회귀 모형이 독립변수와 종속변수 사이의 비선형 관계를 나타내지 못해 예측력이 떨어진다는 단점을 보완하기 위해 도입한 모델이다. 가법적인(additive) 구조 안에 비선형 함수를 적합시킬 수 있어 학습 성능과 설명 가능성을 동시에 달성한다.

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \rightarrow y = \beta_0 + f_1(X_1) + ... + f_n(X_n)$$

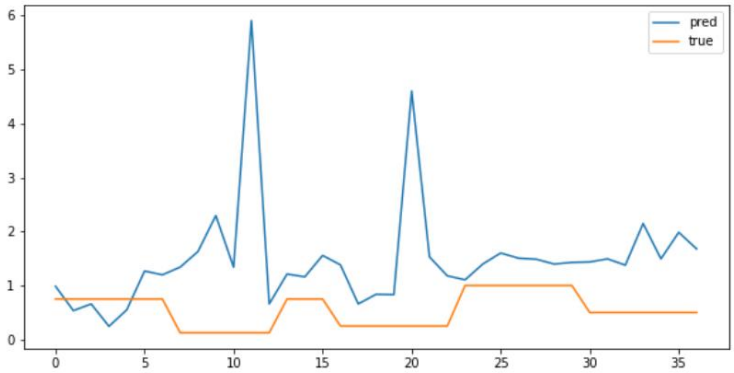
Pygam 패키지의 LinearGAM 은 대표적으로 2 개의 하이퍼 파라미터를 입력 받는다. 첫번째 값으로, *n_splines 가 커질수록 더 유연한 모델을 fitting 할 수 있지만 과적합 문제가 생길 수 있고 독립변수와 종속변수 사이 관계를 해석하기 어렵다*는 단점이 커진다. 따라서 본 모델에서는 범위를 4 에서 10 사이로 제한하였다. *lambda 는 비선형 함수의 유연성을 조절*하는 역할을 한다. 작은 lambda 값은 더 부드러운 함수를 생성하고, 큰 lambda 값은 더 직선에 가까운 함수를 생성하므로 적절한 lambda 값을 통해 과적합을 방지할 필요가 있다.

본 팀은 구축한 데이터셋을 평가하기 위해 **(1) 시기변수를 도입하기 전 데이터, (2) EDA 방식으로 시기변수를 도입한 데이터, (3) UMAP 방식으로 시기변수를 도입한 데이터** 총 3 가지에 대해 순차적으로 GAM 적합을 시도하였다. 하이퍼파라미터 튜닝은 **optuna 방식과 grid-search 방식**을 사용하였으며, 결과적으로 시기변수 도입 데이터가 그렇지 않은 데이터보다, UMAP 방식이 EDA 방식보다, grid-search 방식이 optuna 방식보다 준수한 예측 성능을 보임이 확인되었다.

나. 모델 학습 및 성능 평가

```
LinearGAM
============================================= ==========================================================
Distribution:                      NormalDist Effective DoF:                                     48.3691
Link Function:                  IdentityLink Log Likelihood:                                -30381.6497
Number of Samples:                     20851 AIC:                                           60862.0376
                                             AICc:                                          60862.2767
                                             GCV:                                               0.2712
                                             Scale:                                             0.27
                                             Pseudo R-Squared:                                  0.7306
==========================================================================================================
Feature Function                  Lambda               Rank         EDoF         P > x        Sig. Code
================================= ==================== ============ ============ ============ ============
s(0)                              [0.02]               8            6.9          1.11e-16     ***
s(1)                              [0.02]               8            6.5          1.11e-16     ***
s(2)                              [0.02]               8            6.9          1.11e-16     ***
s(3)                              [0.02]               8            6.6          1.11e-16     ***
s(4)                              [0.02]               8            6.8          1.11e-16     ***
s(5)                              [0.02]               8            6.8          1.11e-16     ***
s(6)                              [0.02]               8            6.9          1.11e-16     ***
f(7)                              [0.02]               2            1.0          1.11e-16     ***
f(8)                              [0.02]               2            0.0          1.11e-16     ***
intercept                                              1            0.0          1.22e-01
==========================================================================================================
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

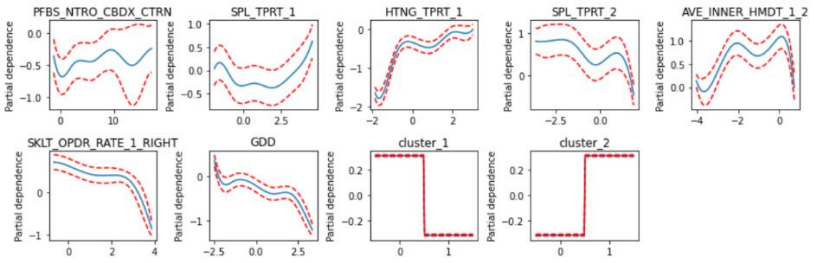
가장 준수한 성능을 보인 UMAP & optuna의 사례를 기준으로, cluster_1, cluster_2에 대해서는 s() 함수 대신, f()함수가 사용되었으며 n_splines = 8, lambda = 0.0200486304871767 97의 값이 선택되었다. 따라서 해당 모델을 통해 최종 예측 결과를 도출하였다.



결과적으로 test set에 대해서는 약 1.474의 아쉬운 RMSE 값이 구해졌고, 이는 **하반기의 급격한 추세 변화를 모델이 효과적으로 반영하고 있지 못함을 의미**한다. GAM이 갖는 설명적 측면의 장점을 고려하여, 해당 모델은 예측력보단 해석력을 중점에 둔 모델로서, **train set에서 나타나는 종속변수와 독립변수 사이 관계를 확인하는 용도로 사용하기로 결정**하였다.

다. 결과 해석

GAM 에서 각 계수에 대한 function 은 plot 을 통해 시각적으로 확인할 수 있다. 아래의 **Partial Dependence Plot 에서 Y 축은 각 변수가 종속 변수에 미치는 영향정도를 나타낸다**. 즉, 확인하고자 하는 종속변수에 대해서 다른변수들을 marginalizing 을 하는 것이다. 이를 바탕으로 연구자는 각 변수와 종속변수 간의 관계성을 분석할 수 있다. 다만 명확한 계수값을 알 수 있는 다중선형회귀와 달리, smoothing function 을 그래프로만 판단해야 해서 객관성이 떨어짐은 감안해야 한다.



시각화 결과를 분석해보자면, 'SKLT_OPDR_RATE_1_RIGHT','GDD'의 경우에 착과수와 *반비례* 관계를 보이며, 'HTNG_TPRT_1'에 경우에는 착과수와 *비례* 하는 관계를 보인다. 또한 범주형 변수에 대해서, cluster_1 에 해당할 때의 착과수가 그렇지 않을 때(cluster_2)의 착과수보다 작았다. 이는 cluster_2 에 해당하는 시기에 착과수가 많았던 경향성을 반영한다.

다만 'PFBS_NTRO_CBDX_CTRN', 'SPL_TPRT_1', 'SPL_TPRT_2','AVE_INNER_HMDT_1_2'의 경우에는 착과수와 관계를 거의 확인할 수 없었다. 이들의 관계를 해석하기 위해서는 전체 작기에 대해서가 아닌 시기 별 분석이 필요할 것으로 판단되었다. 따라서 이 경우, **piecewise regression 의 사후 분석을 통해 관계를 더 확인해볼 수 있을 것**이다.

다만 이와 같은 비례 및 반비례 관계를 해석할 때는 이들의 관계가 **인과관계가 아닌 상관관계**에 불과함을 인지해야 한다. 예를 들어, 천창의 열림 정도가 작았던 시기에 착과수가 크게 기록된 것은 맞지만 이것이 단순히 시기에 따른 (즉 착과수가 시기별로 달라지는 성질에 의한) 상관인지, 아니면 천창의 열림 정도를 작게 조절하는 것이 착과수 증가에 도움이 되는지는 판단할 수 없다. 따라서 **단순한 생육 관리 자동화를 넘어 생육 변수 증대를 위한 최적 환경을 조성하는 것을 목적으로 한다면, 인과관계를 반영하는 전혀 다른 패러다임의 도입이 필요**할 것으로 보인다.

3. Auto ML 모형

가. 사용 모델

Pycaret 은 강력한 Auto ML 라이브러리로, Scikit-learn 패키지를 기반으로 하여 Classification, Regression, Clustering 등의 모델을 지원한다. Pycaret 을 통해 간단한 코드로도 다양한 머신러닝 모델을 한 번에 학습하여 결과를 비교해볼 수 있으며, 하이퍼 파라미터 최적화 및 모델 앙상블 역시 수행 가능하다. **본 팀은 Pycaret 을 활용하여 LightGBM, XGBoost 등 머신러닝 기반 boosting 모델들을 적합하고 그들의 성능을 비교하여 ML 기반의 최종 모델을 결정**하고자 하였다.

모델은 시기 구분선의 기준(UMAP 기반과 EDA 기반)과 Train 과정에서의 교차 검증 방식(K-fold 와 Time Series Cross Validation)에 따라 크게 4 개의 set 으로 나누어진다. 각 set 에 대해, Pycaret 에서 제공하는 기본 모델을 *RMSE 를 기준으로 Sorting 했을 때의 상위 7 개 모델을 개별적으로 train* 하였다. 예를 들어 UMAP + K Fold 조건에서는 Extra Tree, Gradient Boosting, LightGBM, Ada, Catboost, Random Forest, XGBoost 의 7 개 모델이 개별 생성되었다.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.5998	0.6926	0.6951	-7.0941	0.3644	2.3815	0.4250
gbr	Gradient Boosting Regressor	0.6464	0.7248	0.7373	-7.7417	0.3720	2.7182	0.5340
lightgbm	Light Gradient Boosting Machine	0.6362	0.7683	0.7444	-8.0743	0.3825	2.3298	0.1370
ada	AdaBoost Regressor	0.6585	0.7834	0.7564	-7.6766	0.3658	2.8392	0.2680
catboost	CatBoost Regressor	0.6522	0.7647	0.7620	-8.0926	0.3780	2.6901	3.3080
rf	Random Forest Regressor	0.6668	0.8807	0.8052	-9.7323	0.3958	2.5422	0.8110
xgboost	Extreme Gradient Boosting	0.7015	0.8863	0.8261	-9.7540	0.4122	2.9444	0.1620
knn	K Neighbors Regressor	0.6933	0.8533	0.8365	-10.7142	0.4199	2.8876	0.0670
dt	Decision Tree Regressor	0.6972	0.9624	0.8700	-10.8690	0.4148	2.6969	0.0330
en	Elastic Net	0.8206	1.1881	0.8870	-11.2149	0.4670	1.1273	0.0250

▲ UMAP & K-Fold 조건에서 파라미터 튜닝 전 RMSE를 기준으로 나열한 모델 성능 순위 (상위 일부)

이때 상위 모델에 linear regression, ridge regression 과 같이 이미 수행된 분석 방법이 포함되어 있을 경우 이는 제외하고 Optuna 를 이용하여 하이퍼 파라미터 튜닝을 진행하였다. 이렇게 구축된 모델을 기반으로, 각 set 에 대해 모델 blending 역시 시도하였다. Blending 은 대표적인 모델 앙상블 기법으로, *(1) Train 성능이 준수한 상위 n 개 모델 혼합, (2) Test 성능이 준수한 상위 n 개 모델 혼합, (3) Train 성능이 가장 좋은 모델과 Test 성능이 가장 좋은 모델 혼합*을 기본적으로 진행하였다.

다만 아쉽게도 4 개의 경우 모두 앙상블이 유의미하지 못했으며, UMAP + K fold 조건에서는 Cat Boosting Regressor, 나머지 조건에서는 Gradient Boosting Regressor 의 단일 모형이 가장 좋은 test 성능을 보였다. 이들을 비교함으로써 **EDA + Time Series CV 데이터를 이용한 Gradient Boosting 모델**을 최종 모델로 선정하였다.

나. 모델 학습 및 성능 평가

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.2772	0.1446	0.3803	-0.5240	0.2411	0.7076
1	0.6816	0.6632	0.8144	-4.6464	0.4903	0.6693
2	0.8758	0.9536	0.9765	-8.3689	0.5007	0.5730
3	0.6498	0.4914	0.7010	-5.8470	0.3679	0.9427
4	0.6243	0.5978	0.7732	-0.3426	0.4122	4.9219
5	0.8063	0.8628	0.9289	-9.6548	0.3096	1.9007
6	0.2935	0.1465	0.3827	-12.9686	0.2128	2.3225
7	0.3563	0.2445	0.4945	-3.8101	0.2439	1.9230
8	0.6312	0.6652	0.8156	-1.4588	0.4628	0.7434
9	1.0214	1.2153	1.1024	-12.7636	0.4307	0.4676
Mean	0.6217	0.5985	0.7369	-6.0385	0.3672	1.5172
Std	0.2366	0.3365	0.2354	4.5003	0.1032	1.2980

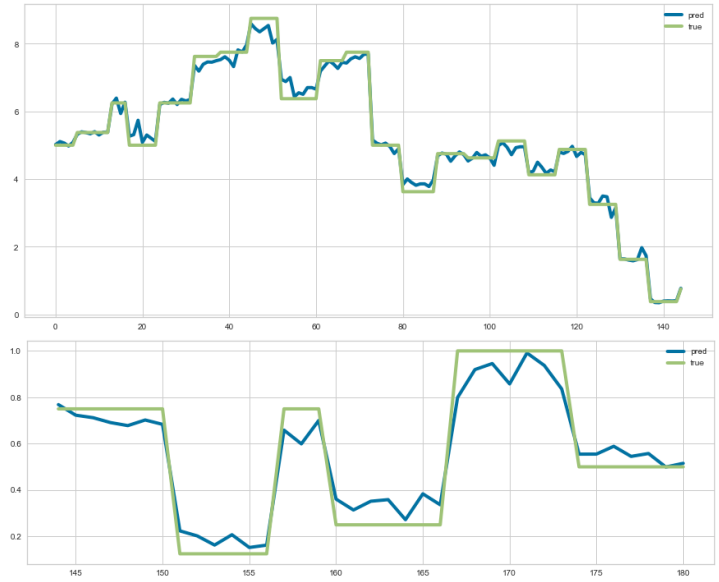
▲ 위는 시계열 데이터에 적용하기 적합한 Time Series Cross Validation 결과이며,
역정규화 전의 수치임에 유의하여 해석해야 한다.

특별히 Cross Validation 시 특정 fold 에서 유독 RMSE 가 크게 관찰되었는데, 이처럼 큰 편차는 UMAP 조건에서 EDA 조건보다 심했다. 따라서 해당 문제는 시계열 데이터의 특성, 또는 단계 구분선의 문제일 것이라고 추정된다.

```
(Pipeline(memory=Memory(location=None),
          steps=[('numerical_imputer',
                  TransformerWrapper(include=['PFBS_NTRO_CBDX_CTRN',
                                              'SPL_TPRT_1', 'HTNG_TPRT_1',
                                              'SPL_TPRT_2', 'AVE_INNER_HMDT_1_2',
                                              'SKLT_OPDR_RATE_1_RIGHT', 'GDD',
                                              'cluster_1', 'cluster_2',
                                              'cluster_3'],
                                     transformer=SimpleImputer())),
                 ('categorical_imputer',
                  TransformerWrapper(include=[],
                                     transformer=SimpleImputer(strategy='most_frequent'))),
                 ('actual_estimator',
                  GradientBoostingRegressor(learning_rate=0.09175312230209909,
                                            max_depth=6,
                                            max_features=0.47181637000768367,
                                            min_impurity_decrease=0.0019632366887038038,
                                            min_samples_leaf=3,
                                            min_samples_split=7,
                                            n_estimators=175, random_state=5952,
                                            subsample=0.3912348610778182))]),
 'Gradient_Boosting_EDA_TS.pkl')
```

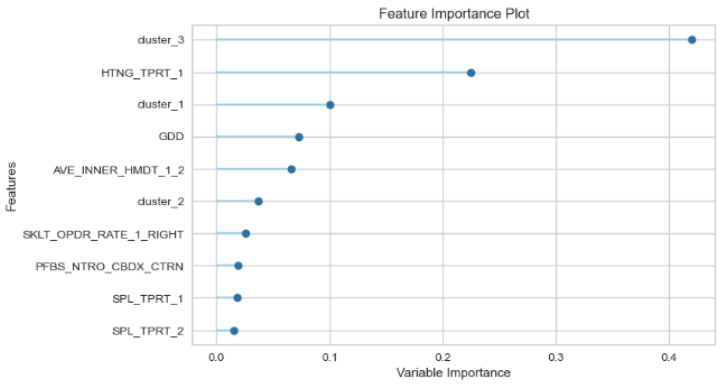
▲ 완성된 Gradient Boosting Model의 하이퍼 파라미터를 위 사진에서 확인할 수 있다.

모델을 구축한 후 다음과 같은 과정을 거쳐주었다. 첫째, 현재의 예측된 y value 는 train data 를 통해 정규화가 되어있는 상태이므로, 오차를 계산할 때 다시 *역정규화*를 시켜주어야 한다. 둘째, *해당 일자의 예측값을 동일 날짜의 모든 time point 착과수에 대한 예측값의 평균으로 대체* 한다. 결과적으로 Train 과 Test 각각에 대해 예측값(파란선)과 실제값(초록선)을 시각화하면 아래와 같다. 시각적으로도 양호한 학습 성능을 확인할 수 있다.

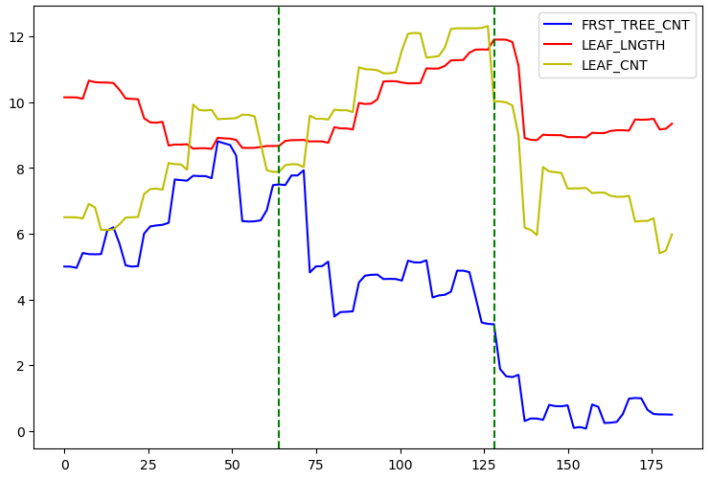


위의 두 그래프는 각각 Train, Test 에 대해 착과수의 실제값과 예측값을 비교한 Time Series Graph 이다. 수치적으로 계산된 Train 의 성능은 **RMSE 0.216, R2 0.989** 이며, Test 성능은 **RMSE 0.084, R2 는 0.926** 으로 상당히 준수했다. Train set 의 RMSE 가 Test set 의 RMSE 보다 크게 계산되었으나, *Test dataset sampling 에 편향이 있었음*을 고려할 때 이를 underfitting 으로 단정짓기는 무리가 있다고 판단하였다. Underfitting 의 경우 Test data 에서도 좋은 성능을 기대하기 어려우며 현재 서로 유사한 데이터만이 Test data 로 분류된 상태이다. 따라서 이러한 차이는 *학습보다는 데이터 자체의 분산에서 기인한 것*으로 이해되었다.

다. 결과 해석



최종 모델을 통해 계산된 Feature Importance 를 시각화한 자료이다. Cluster 3 이 존재하지 않는 UMAP 조건과 달리 EDA 조건에서는 **좋은 성능을 보인 모델들이 공통적으로 Cluster 3 을 매우 중요한 변수로 간주**하였다. 이에 대한 해석을 위해 LEAF_LNGTH 와 LEAF_CNT 의 EDA 를 통해 제작한 시기 변수 Plot 에 FRST_TREE_CNT 를 중첩해서 그려보았고, Cluster 3 이 다른 Cluster 에 비해 전체적으로 낮은 착과수를 보였음이 확인되었다.

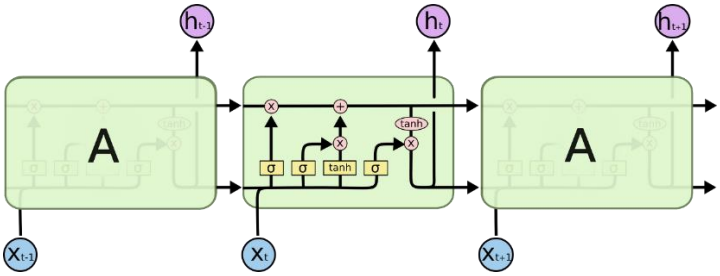


이 외 HTNG_TPRT_1 와 같이 **내부의 온도 및 습도에 영향을 미칠 수 있는 동시에 시기와 강한 상관을 지닌 변수**가 주요 변수로 발견되었다. 이러한 ML 기반 모델들 역시 기본적으로는 블랙박스 모델이기 때문에, Feature Importance 그 이상의 해석을 내리기에는 무리가 있었다. 다만 GAM, Piecewise Regression 과 같은 설명력 있는 모델을 함께 train 시키고 결과를 통합하여 해석함으로써 이러한 문제를 보완하고자 하였다. 관련된 내용은 뒤 섹션에서 자세하게 논의될 것이다.

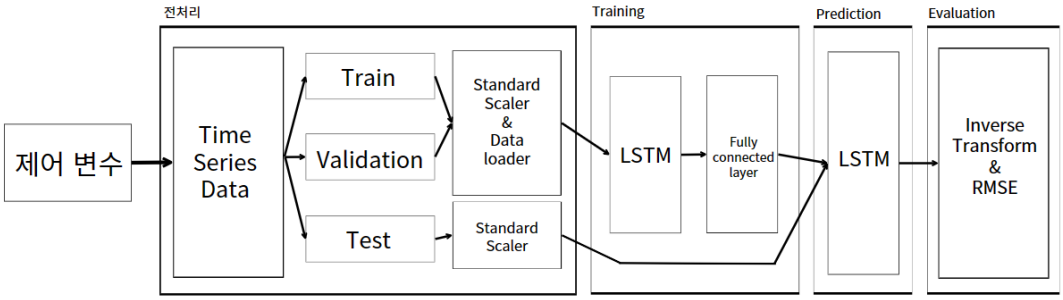
4. LSTM 모형

가. 사용 모델

제공된 데이터가 **시계열 자료**임을 고려하여 딥러닝 모델 중 시간의 흐름을 반영할 수 있는 시계열 모델인 LSTM 을 선택하였다. LSTM 은 RNN 의 발전된 모델로서, 기존의 RNN 모델이 가지고 있던 장기 의존성 문제를 고안하고자 제시된 모델이다.



LSTM은 state에서 단기 상태를 나타내는 *ht*와 장기 상태를 나타내는 *ct*로 나누어진다. LSTM은 3개의 gate를 통해 장기 상태에서 기억할 부분, 삭제할 부분, 읽을 부분을 학습한다. Forget gate를 지나면서 일부는 정보를 읽고, 그 다음 input gate로부터 새로운 정보 일부를 추가하여 타임 스텝마다 일부의 기억을 삭제 및 추가하는 과정을 거친다. 그리고 덧셈 연산 이후 output gate의 tanh 함수로 전달되어 *yt* 를 도출한다. **본 팀이 구축한 LSTM의 모델 구조는 아래와 같다.**

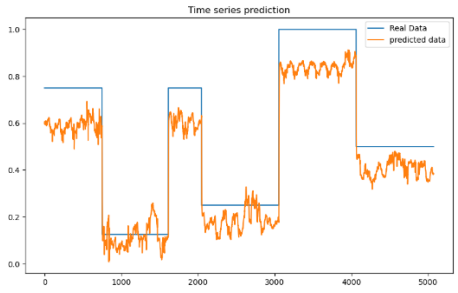


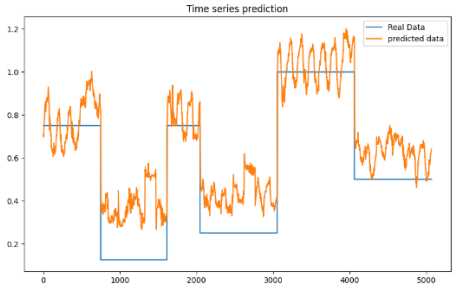
선택된 제어 변수를 바탕으로 해당 시계열 데이터를 Train 데이터셋, Validation 데이터셋, Test 데이터셋으로 나누고, 정규화 및 LSTM 모델에 맞게 데이터 구조를 변형하는 전처리 과정을 거친다. 다음으로 LSTM 모델을 구성하고, 하이퍼 파라미터 튜닝을 통해 도출된 파라미터를 활용하여 모델을 훈련시킨다. 이후 Test 데이터를 활용하여 해당 모델이 어느 정도의 예측 성능을 내는지 RMSE 지표를 활용하여 평가한다.

나. 모델 학습 및 성능 평가

LSTM 모델에 적합한 데이터 형태를 구축하기 위해서, 주어진 데이터를 sliding window 방식을 통해 sequence 데이터로 생성하는 과정을 거쳤다. Sequence 길이는 하이퍼파라미터 튜닝 과정을 통해 144개(하루)로 설정하였다. 특별히 LSTM에 대해서는 시기 변수를 사용하지 않아도 시간적 구조를 학습에 반영할 수 있을 것이라 판단하여, 선택된 7개의 초기 변수와 착과수로 이루어진 데이터를 사용하였다.

아래는 하이퍼 파라미터 튜닝 결과와 무관하게 초기에 돌린 임시 모델로, 임의의 파라미터 값을 설정했음에도 나쁘지 않은 예측 성능을 보이는 것을 확인할 수 있었다. 나머지 파라미터를 고정한 채 hidden layer의 수만 10개, 5개로 다르게 지정했을 때의 결과이다. Training 과정에서의 loss값도 잘 줄어드는 편이었고, test set에 대한 예측 성능을 RMSE로 판단하였을 때 약 0.168, 0.123 정도로 잘 예측됨이 확인되었다.

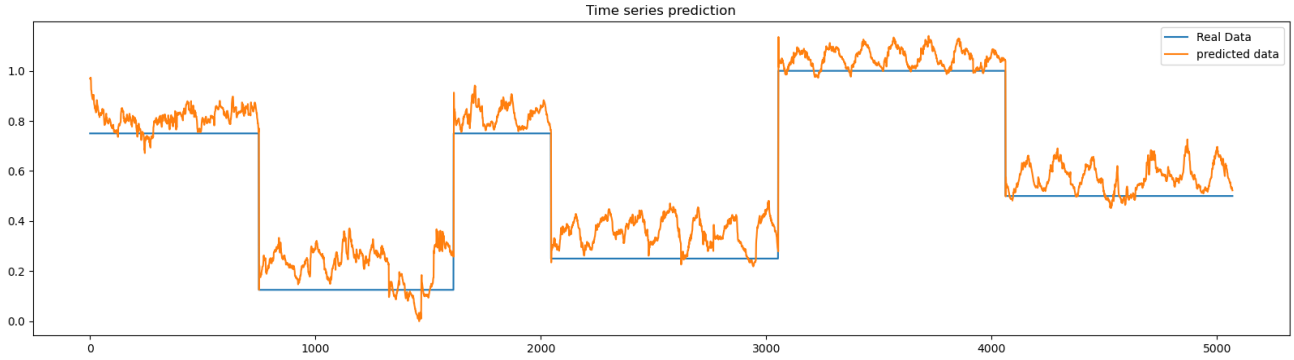




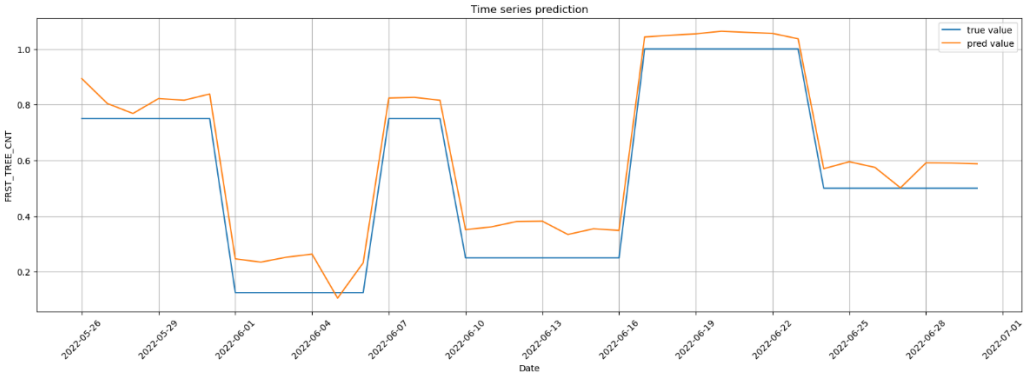
hidden_dim	[64, 128, 256]
seq_length	[36, 72, 144]
layers	[1, 2, 3]
lr	[0.001, 0.01, 0.1]

hidden_dim	256
seq_length	144
layers	1
lr	0.001

하이퍼 파라미터 튜닝에 대해서는 (LSTM 모델의 복잡성과 적절한 튜닝 시간을 고려하여) 지정한 범위 내에서 모든 경우의 조합을 탐색하며 진행하는 'Grid Search'를 이용하였다. 아래의 표 두개 중 왼쪽이 서치를 진행할 후보값이었고, 오른쪽이 최종적으로 설정한 값이다. 파라미터 튜닝 결과, hidden layer를 256으로 설정하여 더 깊은 모델을 구성함으로써 이전의 LSTM 모델보다 좋은 예측 성능을 보일 수 있었다.



RMSE는 약 0.099 정도였으며, plot으로 시각화 하였을 때 이전의 plot들 보다 실제 값과 예측 값 사이의 차이가 줄어든 것을 확인할 수 있었다. 다른 모델들과 마찬가지로 예측된 하루의 착과수 값 평균을 최종 예측값으로 사용하였다. 아래가 시각화 결과이며, 이 경우의 **RMSE는 약 0.088 정도로** 앞선 과정들의 RMSE 값보다 더욱 개선되었다.



다. 한계점

본 팀이 구축한 모델은 test data를 상당히 잘 예측하고 있다. *다만 블랙박스 모델인 LSTM의 특성 상 모델의 예측에 어떤 변수가 어떻게 영향을 주었는지를 파악하기 어렵다.* 이에 대한 보완은 다음 섹션에서 이루어질 예정이다.

더불어 LSTM은 하이퍼 파라미터와 Sequence 길이 등 연구자가 직접 설정해야하는 변수 값의 영향을 많이 받는다. 환경적 제약으로 인해 더 넓은 범위에서 적절한 하이퍼 파라미터를 찾지 못한 것이 모델 성능 향상을 방해하는 요인이 되었을 수 있다. 다른 머신러닝 & 딥러닝 모델에 비해 LSTM은 데이터 자체의 특징에서부터 모델이 달라지기 때문에, 데이터 자체에 따른 변동성이 큰 모델이라고 할 수 있다.

LSTM은 데이터의 시간적 구조를 학습하는 가장 대표적인 모델이지만, 최근 관련 분야의 발전이 이루어지며 시계열 데이터를 반영할 수 있는 딥러닝 모델의 다양성이 증가하였다. 그 중 빠질 수 없는 것이 Transformer이다. 추후 Transformer 기반 autofomer, DLinear와 같은 최신 모델을 사용하여 성능을 확인하고, 기존의 ML 및 LSTM 모델과 비교할 수 있다면 예측 모델 개발에 있어 큰 도움이 될 것이라 사료된다.

5. 통합적 결과 해석

가. 모델링 결과

본 섹션에서는 GAM 을 비롯한 회귀 모형 적합 결과와 Gradient Boosting 결과를 통합적으로 해석하고자 한다. 우선 GAM 시각화 결과 'SKLT_OPDR_RATE_1_RIGHT','GDD'의 경우에 착과수와 *반비례*, 'HTNG_TPRT_1'는 *비례*하는 경향을 확인할 수 있었다. 다중선형회귀(MLR) 적합을 통해 회귀 계수를 확인했을 때 HTNG_TPRT_1 의 회귀계수가 0.4063, SKLT_OPDR_RATE_1_RIGHT 의 회귀계수가 -0.2667 로, GAM 의 결과와 일치하는 유의미한 비례 및 반비례 관계를 보였다.

이때 온실 내에서 측정된 난방온도, 일광과 환시를 조절하는데 사용되는 천창의 열림 정도, 생장에 필요한 온도조건은 공통적으로 '**온도**'와 관련된 변수이다. 특히 **HTNG_TPRT_1** 과 **GDD** 의 경우 Gradient Boosting 에서도 높은 Feature Importance 를 기록하여, 두 모델이 어느정도 유사한 판단 기준을 공유함을 유추할 수 있었다.

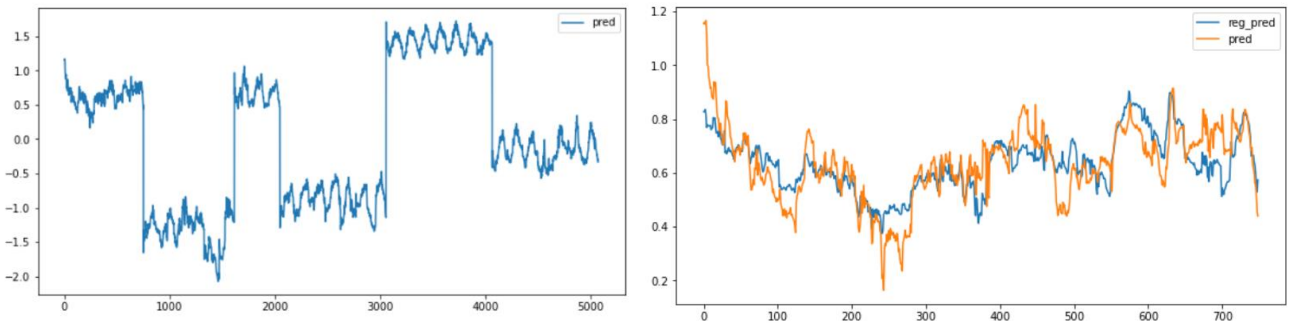
나. 모델 학습 및 성능 평가

XAI 는 설명 가능한 인공지능의 약자로, 블랙박스 모델의 예측 근거에 대한 합리적 설명을 내리기 위해 연구되고 있다. 본 팀은 원래 LRP, LIME, SHAP 과 같은 XAI 모델을 통해 LSTM 모델을 설명하고자 했었다. LRP 를 통해 분석에 사용된 각 feature 가 최종 output 을 도출하는데 발휘한 기여도를 스코어링하고, LIME 을 이용하여 예측이 잘 된, 또는 그렇지 않은 국소 cluster 에 회귀 모델을 적합하여 예측된 결과를 설명하는 것이 본래의 목표였다.

그러나 **LRP** 의 경우 다른 XAI 기법들보다 활용 사례를 찾는 것이 쉽지 않았고, 특히 기존의 레퍼런스를 우리의 Uni-directional LSTM 모델에 적용하기 어려움을 코드 분석을 통해 알게 되었다. 또한 **LIME** 은 explanation 을 생성하기 위해서 예측 확률을 계산하는데, 본 팀의 LSTM 모델은 Recurrent 모델을 기반으로 (분류에 대한 확률값이 아닌) Regression 을 수행하므로 도입이 부적절하다고 판단하였다. 실제로 LIME 레퍼런스를 시계열 데이터 예측 문제에 적용시켰을 때 다음과 같은 에러가 발생함을 확인하였다.

```
NotImplementedError: LIME does not currently support classifier models without probability scores. If this conflicts with your use c
ase, please let us know: https://github.com/datascienceinc/lime/issues/16
```

이에 대한 대안으로서 또 다른 feature 기여도 수치인 **SHAP** 를 생각했으나, SHAP 의 경우 pytorch 기반의 LSTM 에 모델에는 지원되지 않음을 확인하였다. 따라서 본 팀은 *LSTM 모델이 왜 특정한 local point 를 해당 값으로 예측했는 지에 대한 해석을 내리기 위해, y 를 LSTM 의 착과수 예측값, x 를 LSTM 이 input 으로 받은 실제 데이터의 내부 변수 값으로 두고 구간 별 회귀식을 적합*하였다. 분석이 목적이므로 train test split 및 일별 평균 처리는 하지 않았다.



▲ LSTM의 예측 결과(좌)와 회귀식 적합의 예시(우)

위는 LSTM 으로 예측한 하반기 20%의 착과수이다. 착과수 예측값이 크게 변화하는 지역을 기준으로 구간을 6 개로 나누어 회귀식을 적합했으며, 결과적으로 GDD 와 SPL_TPRT1 은 전반적으로 양의 회귀 계수를, AVE_INNER_HMDT_1_2 와 PFBS_NTRO_CBDX_CTRN 은 음의 회귀 계수를 보였다. 이와 같은 모델 적합 결과의 예시는 아래와 같다.

$$y = 0.6559 - 0.0559x_1 - 0.0248x_2 - 0.0636x_4 - 0.0403x_5 - 0.04425x_6 - 0.06011x_7$$

이때 6 개의 회귀식에서 공통적으로 HTNG_TPRT_1 의 회귀 계수가 0 이었다. 그 이유는 **6 월 하반에 해당 농가에서 난방기를 사용하지 않아 난방 온도가 0 으로 잡혔기 때문**이었다 (어떻게 보면 HTNG_TPRT_1 이 0 의 값을 갖기 때문에 양의 회귀계수가 곱해졌음에도 값의 변화가 없어 착과수가 낮은 값을 유지했다고도 해석할 수 있다). *즉 다른 시기의 데이터에 대해서 회귀식 적합을 했다면 GAM, Gradient Boosting 과 마찬가지로 HTNG_TPRT_1 이 중요 변수로 파악되었을 수 있지만, 단편적으로 6 월 하반의 데이터만을 고려하면 이러한 사실이 반영될 수 없다.* 나아가 이전의 예측값을 이후 값 예측에 사용하는 LSTM 의 특성과 데이터 내에 존재하는 비선형적 관계는 회귀식에 반영하기 어렵다.

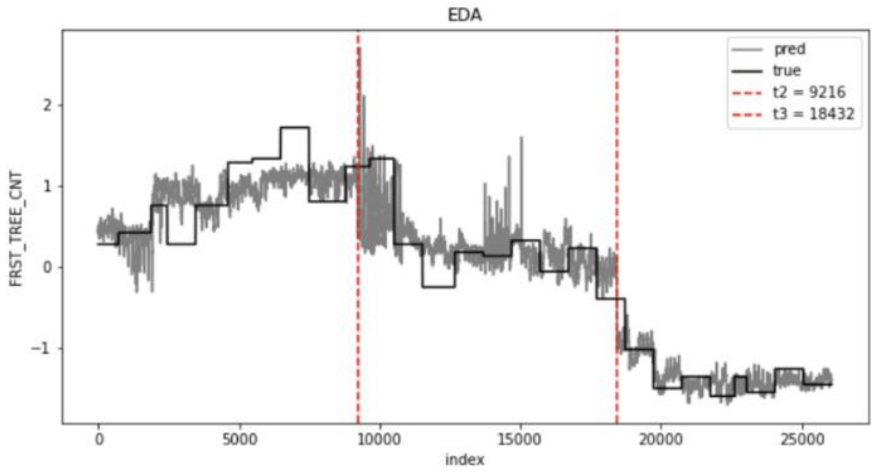
따라서 이와 같은 해석 방법은 **시도에 의의를 둘 뿐, 온전한 접근 방식이 아님**에 유의해야 한다. XAI 알고리즘이 더 많이 연구되고 상용화된다면 각종 농작물의 생육 예측 모델을 해석하는 데 있어서도 큰 발전을 꾀할 수 있을 것으로 기대된다.

IV. 사후 결과 분석

1. 설명 변수 추가 분석

가. 부분적 회귀 적합 (Piecewise Regression)

앞서 GAM 과 MLR 로 종속변수에 대한 독립변수의 영향력을 확인을 하는 분석을 진행했었다. 그러나 전체 작기에 대해서 착과수와 환경변수 사이의 회귀모델 혹은 GAM 을 적합시킬 때, 'PFBS_NTRO_CBDX_CTRN', 'SPL_TPRT_1', 'SPL_TPRT_2','AVE_INNER_HMDT_1_2'의 경우에는 착과수와 관계 파악을 위한 전체적인 추세를 확인하기가 어려움이 있었다. 따라서 전체 작기에 대해서 분석을 진행하는 것이 아니라, 단계별로 데이터를 분리해서 각 단계별 착과수와 독립변수와의 관계를 확인하는 분석을 추가적으로 시도하는 **Piecewise regression** 을 진행하였다. 가장 성능이 좋았던 예측 모델인 Gradient Boosting 이 EDA 를 이용해 시기 변수를 도입한 데이터셋을 이용했으므로, 이를 중심으로 분석을 시행하고 결과를 해석하였다.



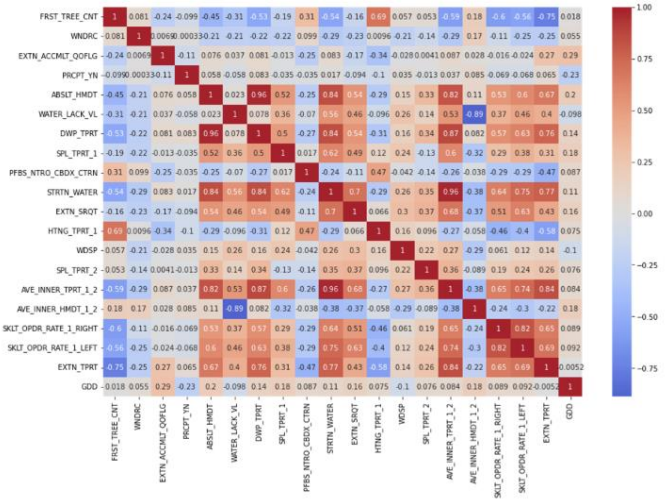
(a) 첫 번째 시기에 해당하는 2022-03-06 0:00 부터 2022-05-08 23:50 에는 PFBS_NTRO_CBDX_CTRN, SPL_TPRT_1, AVE_INNER_HMDT_1_2 와 반비례, SPL_TPRT_2 와 비례 관계가 나타났다. 또한 이 시기에는 **GDD** 가 가장 중요도가 가장 높게 나타났다.

(b) 두 번째 시기에 해당하는 2022-03-06 0:00 부터 2022-05-08 23:50 에는 PFBS_NTRO_CBDX_CTRN, SPL_TPRT_1, AVE_INNER_HMDT_1_2, SPL_TPRT_2 와 반비례 관계가 나타났고, 또한 이 시기에는 **AVE_INNER_HMDT_1_2** 가 가장 중요도가 크게 나왔다.

(c) 마지막 시기에 해당하는 2022-05-09 0:00 부터 2022-06-30 23:50 의 시기에는 PFBS_NTRO_CBDX_CTRN 와 반비례, SPL_TPRT_1, AVE_INNER_HMDT_1_2, SPL_TPRT_2 와 비례 관계가 나타났고, 또한 이 시기에는 **PFBS_NTRO_CBDX_CTRN** 의 중요도가 가장 크게 나온 것을 확인할 수 있었다.

이러한 시기별 회귀 적합을 통해 단계별로 중요한 환경 변수를 파악할 수 있었고, 전체 작기를 기준으로 모델을 적합했을 때 해석하기 애매했던 환경변수와 착과수 간의 관계를 세밀하게 검증할 수 있었다.

가. 환경 변수 및 제어 변수 간 관계 설명



변수 간의 관계를 추가적으로 설명하기 위해 위와 같이 선택된 변수, 추가적인 자연변수, 그리고 착과수와의 상관관계를 도출했다. 상관관계를 기반으로 외부 변수, 착과수, 내부 변수 등과의 관계를 분석하되, 사전 조사에 의해 필요하다고 생각하여 넣은 변수(GDD)와 추가적으로 중요하다 판단하여 도입된 변수에 대해서는 상관관계에 의존하지 않고 사전 지식을 참고하여 분석하였다. 이를 통해 특정 외부 변수 조성 시기에 조절에 주의해야 할 제어 변수를 규명할 수 있었다.

먼저 PFBS_NTRO_CBDX_CTRN 의 경우 관측지점실내이산화탄소농도와 착과수는 수치상 뚜렷한 상관관계는 아니지만, 양의 상관관계를 나타낸다. 다른 자연 변수와의 관계는 거의 0 에 가까운 수치로 유의미한 상관관계를 보이지 않았으나, 외부 온도를 나타내는 EXTN_TPRT 과의 상관관계는 -0.47 로 나타났다. 즉, 관측지점실내이산화탄소농도와 외부온도는 음의 상관관계를 보임을 알 수 있다. 결론적으로 착과수를 높이기 위해서는 관측지점실내이산화탄소농도는 높아야 하고, 외부 온도는 낮아야 함을 알 수 있다. 즉, **스마트팜 외부의 서늘한 날씨에서도 딸기를 재배하기 위해서는 적절한 스마트팜 내부 온도를 외부 온도의 변화에 맞추어 온도 관련 제어 변수로 조절할 필요가 있다.**

다음으로 SPL_TPRT_1 / SPL_TPRT_2 의 경우 해당 변수도 제어 변수에 속하기 때문에, 해당 변수와 관련이 있으면서 착과수와 유의미한 관계를 가지는 자연 변수를 파악하는 것이 중요하다고 판단하였다. 공급온도_1 은 절대 습도와의 상관관계는 0.52, 외부 일사량과의 상관관계는 0.49 로 나타났고, 공급온도_2 는 절대 습도와의 상관관계는 0.33, 외부 일사량과의 상관관계는 0.37 로 나타났다. 따라서 **절대 습도와 외부 일사량의 변화에 맞추어 스마트팜 내부의 딸기 생육 환경을 공급온도를 통해 적절히 유지하는 것이 중요하다 판단하였다.**

HTNG_TPRT_1 은 난방 온도와 착과수는 0.69 의 다소 강한 양의 상관관계를 보인다. 난방 온도와 자연 변수와의 관계는 EXTN_TPRT 인 외부온도와는 -0.58 정도의 음의 상관관계를 보였다. 적절한 딸기 생육 환경을 조성하기 위해서는 외부 온도가 서늘한 편이 선호되는데, 선호되는 다소 서늘한 기온을 넘어서 **딸기가 제대로 자랄 수 없는 추운 시기에는 이를 조절해줄 적절한 난방 온도와 같은 제어 변수가 필요할 것으로 예상된다.**

AVE_INNER_HMDT_1_2 평균 내부 습도의 경우, 착과수와의 상관관계는 낮지만 사전 조사를 바탕으로 습도와 생육 환경은 밀접한 연관이 있다고 판단하여 추가적으로 넣게 되었다. 평균 내부 습도와 EXTN_SRQT 인 외부 일사량은 -0.37 의 음의 상관관계를 보인다. 따라서 적절한 딸기 생육 습도를 유지하기 위해서는 순환펌프작동여부, 분무장치, 3 방밸브개도율, 천창개도율_1_좌, 천창개도율_1_우와 같은 제어변수를 조절하면 된다. **외부 일사량이 높으면 평균 내부 습도가 낮기 때문에 3 방밸브개도율, 수평스크린개도율, 천창개도율_1_좌, 천창개도율_1_우와 같은 외부 일사량을 제어해주는 변수를 적절하게 사용할 필요가 있다.**

SKLT_OPDR_RATE_1_RIGHT 은 천창개도율_1_우와 착과수는 상관관계가 -0.6 으로, 유의미한 상관관계를 보였다. 여기서는 **이슬점 온도, 절대 습도, 외부 일사량을 제어할 수 있는 변수가 천창개도율_1_우**이다. 이처럼 해당 제어변수를 통해 어떤 자연 변수를 조절할 수 있는 지를 파악하는 것이 최적의 딸기 생육 환경 조성을 위해 중요하다 판단하였다.

GDD 는 다른 자연변수와의 상관관계가 높지 않았으나, 착과수를 예측하는데 있어서 사전 조사에 의해 필요하다 판단되어 새롭게 추가한 변수이다. GDD 는 작물의 발아부터 성숙까지 생육단계에 따라 일정량의 '열량'을 얻어야 성숙된다는 이론을 기반으로 하였다. 따라서 해당 데이터를 기반으로 구한 GDD 변수는 그 시기에 작물이 필요한 열량을 나타낸다고 할 수 있다. 나아가 GDD 를 통해서 해당 계절 및 환경 온도 등을 예측할 수 있기 때문에, 해당 시기의 GDD 를 바탕으로 적절한 딸기 생육 환경을 조성해 주는 온도 관련 제어 변수의 파악이 중요할 것이라 판단하였다. **해당 착과수 데이터의 온도 관련 제어 변수는 냉방기작동여부, 공급온도, 난방온도, 배기온도, 3 방밸브개도율, 이산화탄소설정값이 있다**.

2. 반응 변수 추가 분석

해당 섹션에서는 착과수를 포함한 '생육 변수'에 대해 어떠한 추가적 논의들이 이루어졌으며, 이러한 고민들에서 파생된 추가 아이디어가 어떤 의미를 갖는지를 검토하고자 한다.

가. 착과수 자기회귀 모형 (Autoregressive Model: AR)

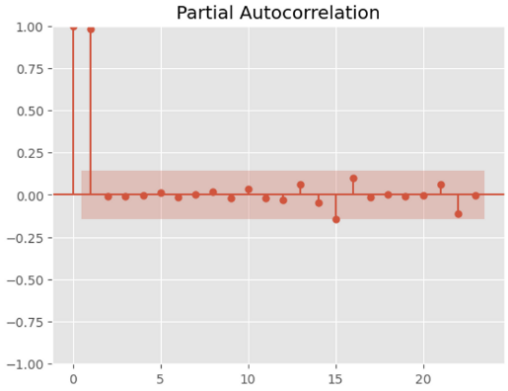
시계열 데이터 분석에 주로 사용되는 대표적 딥러닝 모델로서 LSTM은 이전의 X, y값과 상호작용하여 기준 시점의 y값을 예측한다. 이러한 모델의 구조로부터 본 팀은 '자기회귀모형'에 대한 아이디어를 얻을 수 있었다. *이전의 착과수 변화 추세에 대한 정보만을 가지고 있을 때, '지금과 같이' (동일한 환경을 유지한다는 의미가 아닌 동일한 정도의 최적성을 유지한다는 의미이다) 재배를 지속한다는 가정 하에 기준일 착과수를 예측할 수 있을 것인가?* 지금까지는 환경변수를 독립변수로, 착과수를 종속변수로 설정하고 모델링을 진행했지만, **착과수 데이터만을 가지고 시계열 데이터를 예측하는 모델 적합**을 할 수도 있다. 정상성을 갖는 시계열 데이터가 있다고 가정할 때 자기회귀모형의 수식은 다음과 같다.

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$$

▲ 정상성을 갖는 시계열 데이터가 있다고 가정할 때 자기회귀모형의 수식은 다음과 같다.

자기회귀모형은 이전 시점의 데이터로부터 미래 시점의 데이터를 예측하기 위한 모델이다. 시계열 데이터는 시간에 따른 패턴, 추세, 계절성 등을 포함하고 있으며, 자기회귀모델은 이러한 데이터를 모델링하는데 유용하다. 즉 과거 정보가 현재 정보와 관련이 있다고 가정하여 과거 정보를 활용하여 모델링을 진행할 수 있으며 다중선형회귀와 같이 시차를 고려하지 못하는 모델을 보완하는 하나의 해결책이 될 수 있는 것이다.

모델 적합을 위해 먼저 현재시점의 착과수를 종속변수로 놓고, 이전 7일 간의 착과수를 각각 독립변수로 두어 자기회귀모형에 적합시켰다. 다만 구축된 자기회귀모형은 성공적이지 못했다. R-squared 값은 0.971로 또한 97%의 매우 높은 설명력을 보였으나 t- one 변수(하루 전 착과수)의 회귀계수 값이 0.9931로 매우 높게 나타났다. 또한 PACF 도표에서 lag1 (하루 전 착과수)에서만 유의 상관계수 값을 보였다. 이는 회귀 모형이 이틀 전부터의 착과수를 기준일 착과수 예측에 유용하게 반영하지 못함을 의미하여, 장기적 예측에의 활용 가능성을 제한한다.



이를 보완하기 위한 아이디어는 다음과 같다. 첫째, **'이전의 착과수'를 다른 방식으로 모델에 반영**하는 것이다. 예를 들어 착과수의 절대적 값이 아닌 변화량을 설명 변수로 사용하거나, 바로 직전의 착과수를 분석에 포함시키지 않고 모형을 적합하거나, 이전 n일의 착과수에 대한 합리적 대표값을 사용해볼 수 있다.

둘째, **데이터의 축적**이다. 자기회귀모형, 차분(differencing), 이동평균(moving average)을 결합한 시계열 모형으로서 ARIMA는 더 복잡한 시계열 패턴을 다룰 때 효과적이다. 만약 1월부터 6월까지의 축적된 데이터(ex. 5개년)가 존재한다면 ARIMA 모형은 그 안의 계절성 및 추세를 파악하여 광범위한 작기의 딸기 착과수를 분석할 수 있다. 이와 같은 데이터 측면에서의 보안은보다 다양한 AutoTS 모형들의 사용 역시 가능하게 할 것이다.

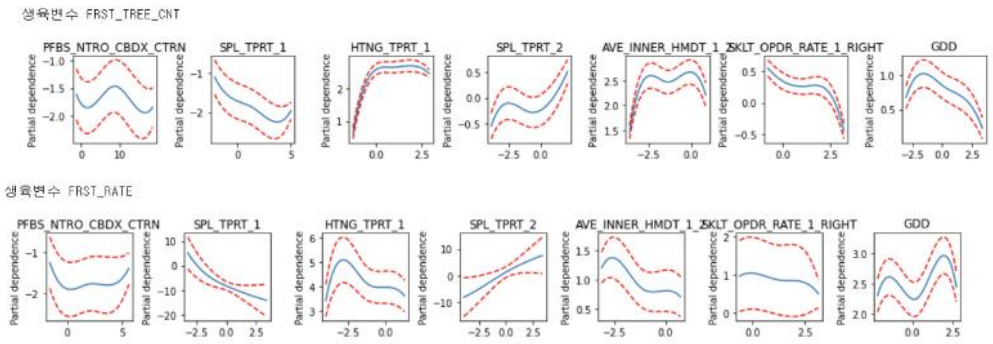
마지막으로는 이전의 데이터를 토대로 이후의 데이터를 예측하는 포맷을 **착과수 뿐만 아닌 생육 데이터 전체로 확장**시켜 볼 수 있다. 이전의 생육 변수 값 (예를 들어 3일 전의 엽폭, 일주일 전의 관부직경 등)

을 설명 변수로 사용하여 기준일 착과수를 예측하는 모형을 적합한다면, 환경 변수를 통한 예측보다 더 '*그럴듯한*' 실시간 예측이 가능해진다.

나. 타 생육 변수 분석

본 팀은 **착과수를 제외한 나머지 생육 변수에 대해서도 회귀 모델을 적용함으로써 타 생육 변수와 착과수 간의 간접적 관계를 파악**하고자 하였다. 해당 단계에서의 분석은 예측을 주된 목적으로 하고 있지 않기 때문에 좋은 설명력을 가진 회귀 모델을 사용하였으며, 착과수와 관련된 해석을 내리기 위해 착과수 예측에 사용된 총 7개의 변수만을 설명 변수로 사용하였다. 총 21개의 데이터셋(csv)을 이용하여 착과수 예측에 사용된 결과변수만을 포함한 데이터 프레임을 구축하였고 결측치를 전부 제거(drop)하였다. 결과적으로 총 21개의 다중선형회귀 모델과 GAM 모델을 적합하여, 회귀 계수 확인 및 변수 간 관계 시각화를 시도하였다.

모든 생육 변수가 동일한 환경 변수로 설명될 것이라는 기대를 하지는 않았기 때문에 크게 유의미한 결과를 얻지는 못했지만, 갖고 있던 도메인 지식을 데이터를 통해 확인할 수는 있었다. 구체적으로, FRST_RATE(착과율)을 종속변수로 하는 분석에서 착과수와 비슷한 경향을 보이는 것을 확인할 수 있었다. 이처럼 특정 종속변수와 관련하여 서로 비슷한 양상을 보이는 생육변수를 확인하는 행위를 통해 우리는 **집중해야할 생육변수의 개수를 줄일 수 있다**.



▲ 공통적으로 SPL_TPRT_1, SKLT_OPKR_RATE_1_RIGHT 감소, HTNG_TPRT_1 증가 추세를 보인다.

스마트팜 농가에서 IoT 기술을 통해 제어 및 관리되는 온실 내, 외부 환경 변수들과 달리, 각종 생육 변수를 상세하게 기록하는 것은 까다로울 수밖에 없다. 그래서인지 생육 변수를 측정한 다양한 데이터들은 매 순간의 변화를 반영하기보다는 일, 주, 보름 단위의 정보만을 담고 있는 경향이 있다. 이때 *비슷한 환경 변수를 이용해 예측이 가능한 생육 변수의 군집을 확인함으로써 측정해야할 주요 생육 변수의 개수를 최소화* 하는 것은 **데이터 수집 측면에서 경제적 의의**를 지닌다.

나아가 이번 분석에서는 시행되지 않았지만, 착과수 예측 모형 구축을 위해 거친 절차를 개별 생육 변수들에 그대로 적용하여 **각 생육 변수를 잘 설명하는 환경 변수 쌍을 규명**하는 것 역시 큰 의미를 가질 것

으로 보인다. 예를 들어 과실 재배에는 '적과'라는 개념이 있는데, 이는 착과수보다 단일 과실의 크기 및 당도 등의 품질을 중시하여, 착과수를 제한함으로써 수확되는 과실 상품성을 향상시키는 것을 목적으로 한다. 정호정 외(2013)의 연구는 적과 그룹과 비적과 그룹을 비교함으로써 딸기 5품종의 최적 착과수를 규명하였고, 노희선과 이윤숙(2020)의 연구는 초장, 엽수, 줄기굵기 등의 생육 변수를 이용하여 토마토 착과수의 변화를 예측하는 모델을 제시하였다.

이러한 선행 연구들은 착과수와 타 생육 변수 간의 상관성을 설명한다는 점에서 의미를 지닌다. 다만 아쉽게도 *타 생육 변수의 최적화에 영향을 미치는 환경 변수 쌍이 착과수 최적화에 요구되는 환경 변수 쌍과 어떻게 유사하거나 다른지를 보여주지 못한다*. 추후 더 자세히 논의되겠지만, 본 팀의 모델과 다양한 선행 연구의 예측 모델이 해결해야할 가장 중요한 과제 중 하나는 나무에 집중하느라 숲을 보지 못하는 것이다. Y 변수들과 X 변수들 간의 총체적인 관계를 모르기 때문에, *온도, 습도, 과중, 당도를 input으로 받았을 때의 착과수를 예측할 수 있을지는 몰라도* **'a 착과수, b 과중, c 당도의 열매를 생산하기 위해 온도 및 습도를 어떻게 조성해야 하는지'는 알 길이 없다.**

이를 보완하는 데에는 생육변수1과 생육변수2 간의 직접적 상관을 확인하는 것보다, 생육변수1을 설명하는 환경변수 모형과 생육변수2를 설명하는 **환경변수 모형 간의 비교분석**을 실시하는 편이 유익할 수 있다. *이는 생육변수들과 환경변수들 간의 다대다 관계를 규명*하기 위한 하나의 아이디어이고, 동일한 목적을 위한 다른 접근 방식이 바로 다음 문단에서 제시될 예정이다.

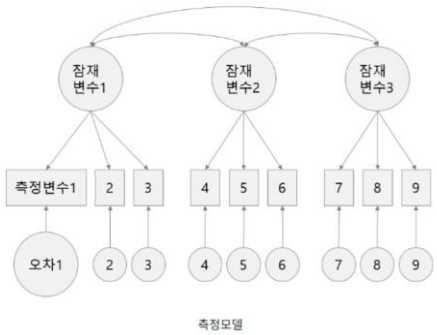
다. 베이지안 방법론을 활용한 구조방정식 모델링

최적의 생육 결과를 얻기 위해서는 온도, 토양, 물, 영양분,비료 , PH & E.C 광합성 등의 조건이 적절히 형성되는 것이 중요하다. 그러나 영양분이 적절하게 공급되고 있는지, 물이 필요에 맞게 적절히 공급되고 있는지 등은 (물론어느정도는 알 수 있겠지만) 직접적으로 측정이 불가능하다. 따라서 **관측된 변수들을 이용하여 이와 같은 잠재변수를 파악할 수 있는 모델**을 제안하고자 한다.

구체적으로 예를 들면, 습도가 지나치게 높으면 잿빛 곰팡이병이 발생할 수 있고, 중심 줄기 옆 작은 곁눈의 발생이 커져 방임시 딸기의 생산량이 크게 줄어들 수 있으며 줄기 길이가 최적수준보다 짧아질 수 있다. 관측가능한 지표들 중 줄기의 길이, 곁눈의 개수, 잿빛 곰팡이병의 발생여부 등을 집중적으로 보면 습도가 적절히 형성되었는지를 더 효율적으로 파악할 수 있을 것이다. 이처럼 *잠재변수가 제대로 형성되지 않았을 때 더 큰 변화를 나타내는 관측변수를 파악해 잠재변수가 식물에 맞게 제대로 형성되었는지 여부를 판단하고자 한다*.

물론 기존 데이터에 포함되어 있는 정보를 이용하여 새로운 변수를 제작함(GDD도 그 예시였다)으로써 잠재변수의 존재를 도입할 수 있다. 그러나 데이터를 분석하는 입장에서 분석가는 *변수 간의 관계에 대한 모든 도메인 지식을 갖고 있을 수는 없으며, 변수 제작 방식에는 주관이 개입*되기 때문에 최대한 '데이터

에서 파생된' 접근 방식이 필요하다. **구조방정식(SEM)**이란, 여러 종류의 변수를 관측하고, 이 변수들을 사용하여 구조를 정의한 다음 이 구조들 간의 관계 모형을 데이터를 이용한 가설 검정을 통해 찾아가는 과정이다. 구조방정식은 회귀분석, 경로분석, 확증적 요인분석의 세 가지 분석 기법이 결합된 분석 방법으로 *잠재변수를 고려할 수 있으며, 여러 변수 간의 영향관계를 동시에 분석할 수 있다*는 장점을 지닌다.



▲ 구조방정식의 경로도. 요인적재값(loading)은 각 잠재변수와 관측변수 간의 상관계수를 의미한다.

구조방정식 절차 중 측정 모델 구성을 위한 과정으로써, 요인분석을 실험적으로 시도해보았다. 네이버 스마트팜 데이터셋에는 '동일 농가에서 동일 시기에 동시 측정한 생육 변수' 데이터가 존재하지 않으므로, 23개의 생육 변수 데이터가 '그럴 것'이라는 강한 가정을 깔고 181일에 대한 값을 합쳐서 하나의 데이터 프레임을 구축하였다. 따라서 이를 통해 구축된 Factor의 의미 해석은 무의미하며, 단순한 과정 구현으로 받아들여야 할 것이다.

	0	1	2
Loadings	6.734768	3.316530	3.256653
Proportion Var	0.292816	0.144197	0.141594
Cumulative Var	0.292816	0.437013	0.578607

구현 결과, 3개의 잠재변수를 사용하였을 때 57%의 정보를 설명할 수 있었으며 각 잠재변수와 유의미한 상관관계를 가지고 있는 생육변수들을 묶어낸 5개의 집합이 구성되었다. 이처럼 생육 변수를 토대로 잠재변수를 추측한 후 그들간의 연관성도 함께 고려하는 구조방정식 모형을 이용한다면, **잠재변수에 대한 관리/감독이 적절히 이루어지고 최적의 딸기 생육을 위한 환경이 잘 조성되고 있는지를 더 효율적으로 판단**할수 있을 것으로 기대된다. 이를 위해서는 데이터의 측면에서 *동일한 환경 하에 측정된 관측 변수(ex:생육지표) 표본이 더 많이 축적되어야 하고, 최적의 데이터 뿐만 아니라 여러 환경에서의 관측변수 데이터가 확보되어야 한다* (뒤따르는 섹션에서 더 자세히 논의될 것이다).

나아가 방법론적 측면에서 **베이지안 접근법**을 도입하는 것이 상당한 기대효과를 지닐 것으로 보인다. 일반적으로 구조방정식모형은 표본의 크기가 충분히 큰(600 이상) 경우 bias(편의)가 없는 해석이 가능하지만 표본의 크기가 작은(300 미만)경우 다변량 정규성가정을 만족시켜도 모집단에 대한 해석에 편의를 배제할 수 없다고 알려져 있다 (Gao 등, 2008). 딸기에 대한 표본 수집에 있어, 농장 수 제약 등으로 600개 이상의 개체(표본)에 대한 정보를 수집하기에는 시간과 비용에 큰 제약이 있을 것이다. 그리고 제안된 베

이지안 방법은 *다양한 사전정보의 사용이 가능하며, 표본의 크기가 작은 경우에도 믿을 만한 결과를 제공한다는 장점이 있다.* 또한 계산된 사후분포를 추후 갱신 자료의 사전분포로 사용할 수 있기에, **스마트팜 데이터가 축적됨에 따라 더욱 정확한 사전정보 반영으로 향상된 모델을 개발하는 것이 가능하다.**

실제로 해당 방법론을 활용해 기후요인과 생산요인 간 관계를 분석한 논문에서는 기후를 나타내는 원본 데이터를 가공해 동일 기후를 나타내는 기간 내의 일평균온도의 평균값, 1월 1일부터 수확까지의 일조시간을 측정한 봄일조시간, 강수량이 1mm이상 기록된 일수의 합인 봄강수일수 등을 모델 관측변수로 두어 잠재변수를 파악하였다. 이를 스마트팜 사례에 적용함으로써 *사전정보를 충분히 반영해 데이터를 가공하고 활용*한다면, 잠재변수를 더욱 뚜렷하게 파악하고 집중할 관측변수에 대한 유의미한 통찰을 얻을 수 있을 것이다.

3. 아이디어 제안

가. 사전 지식 반영을 위한 최적화 모델 제안

<딸기 촉성재배 시 생육 단계별 온도관리 기준>

생육 단계	낮(℃)	밤(℃)	비 고
생육촉진기	28~30	10~13	◦보온 개시 초기는 액화방 분화 시기이므로 낮 30℃, 밤 13℃ 이상 되지 않도록 유의 ◦고설식 수경재배는 밤 온도를 8℃ 이상 유지함
출뢰기	25~26	8~10	
개화기	23~25	5~8	
과실비대기	20~23	5~7	
수확기	20~23	5	

위는 딸기 생육에 있어서 알려져 있는 환경 변수에 대한 도메인 지식이다. 현재의 예측 모델은 이러한 '제약 조건'을 반영하고 있지 못하므로, 이상 온도를 input 으로 받았을 때도 (어쩌면 오히려 정상적인 상황보다) 양호한 착과수를 output 으로 반환할 수 있다. 즉 이러한 사전적 지식을 모델링에 반영할 수 있어야 실생활에 대한 적용 가능성이 향상될 수 있으며, 그 방법으로 **제약 조건을 반영하는 '페널티' 항을 손실 함수에 도입**하는 방식을 제안한다.

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array} \tag{4.1}$$

모델 구축에 사용되는 최적화 도구는 기본적으로 이러한 페널티 항을 포함한 전체 손실을 최소화하도록 한다. 손실함수에 제약조건을 추가하는 방식은 아래와 같이 나타낼 수 있다. 제약조건을 추가한 후, 라그랑지 승수법을 사용해, 손실함수를 재정의하는 방식으로 최적화문제를 만들 수 있다.

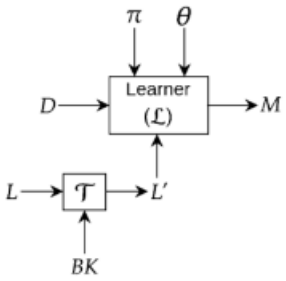


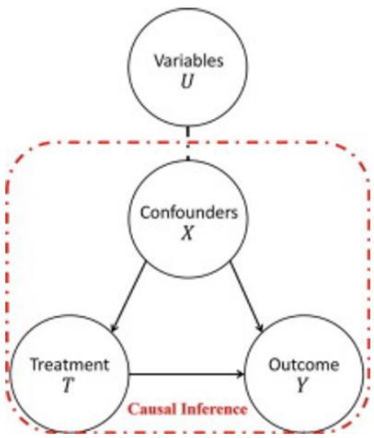
Figure 7. Introducing background knowledge into deep network by transforming the loss function L. $\mathcal{T}$ block takes an input loss L and outputs a new loss function L' by transforming (augmenting or modifying) L based on background knowledge (BK). The learner $\mathcal{L}$ then constructs a deep model using the original data D and the new loss function L'.

제약조건을 손실함수에 다양한 방식으로 추가할 수 있는데, 참고한 논문에 따르면 제약조건을 손실함수에 도입하는 방법으로 Semantic loss'를 들고 있다. 이는 모델의 예측이 도메인 제약 조건을 얼마나 만족시키는 지에 대한 정도에 대한 페널티 항을 도입하는 것이다. 또한 이 논문에서는 손실함수에 페널티항을 추가한 다양한 사례를 소개하고 있다. 그 중 하나로 규칙 기반 도메인 지식을 심층 LSTM 기반 RNN 에 세 가지 다른 수준으로 도입하고, 도메인별 규칙에 기반한 특정 형태의 근사화 및 출력 범위를 디자인하는 것으로 입력을 제한하고, 손실 함수에 페널티 항을 도입한 예시가 있다.

그러나 도메인 지식을 논리적 또는 수치적 제약 조건으로 인코딩하여 심층 네트워크에 통합하는 것은 종종 다루기 어려운 최적화 문제를 해결해야 하거나, 대규모 데이터 셋에 확장가능하지 않다는 문제가 있어 현재 훈련 및 구현 관점에서 몇 가지 어려움을 겪고 있으며, 이러한 어려움을 극복하기 위한 연구가 현재 진행 중에 있다.

나. 인과추론적 관점의 도입 필요성

모델 구축에 있어서 변수 선택에서부터 결과 해석까지의 총체적 과정에 상관분석이 사용되었다. 그러나 농업 의사결정에 있어서 상관관계보다 더욱 중요한 것은 **인과관계**이다. '온도가 a 일 때 착과수가 x 이다'라는 정보는 의사결정에 반영하기에는 불완전하다. **'온도가 a 에서 b 로 변화될 때 착과수가 x 에서 y 로 변화된다'**라는 분석을 할 수 있어야만 더욱 세밀한 환경 조성 가이드라인을 제시할 수 있다.



이러한 관점에서 각종 문제를 다루는 학문 분야가 '**인과 추론(causal inference)**'이다. 인과 추론을 통해 우리는 외생변수가 통제된 상황에서의 가정 사항(what if)에 대한 예측되는 확률적 해답을 내릴 수 있고, 이를 이용하여 의사결정을 내릴 수 있다. 현재 의학 및 사회과학 등의 다양한 분야에서 인과추론을 이용한 의사결정 연구를 진행하고 있으며, 스마트팜 산업에도 각종 제어 장치 조절(개도율, 작동 여부 등)에 있어 이와 같은 관점의 연구 도입을 제안하는 바이다.

V. 결론 및 제언

1. 분석 결과 요약

본 팀은 스마트팜 데이터 플랫폼을 이용한 생산력 증대 및 농업 기술력 발전에 기여하는 것을 목적으로, 네이버 스마트팜 빅데이터 플랫폼에서 제공하는 '**딸기 착과수 최적 환경 데이터' 내의 내부 변수를 이용하여 최적 환경 하에서의 착과수를 예측하는 모델을 학습**하였다. *상관 분석, 라쏘 회귀, 랜덤 포레스트를 통해 본 팀의 목적에 부합하는 변수를 선택하였고, 도메인 지식과 관련 데이터 EDA, 군집화 분석, 차원 축소를 통해 '온도' 및 '시간'을 반영할 수 있는 새로운 변수를 도입* 하였다.

분석에 사용한 모델은 GAM, Pycaret, LSTM 으로, 전역적 및 지역적 회귀 모형 적합을 통해 최적 환경에서의 착과수와 내부 변수 간 상관에 대한 설명력을 확보하고, 다양한 ML/DL 모형 중 최종적으로 가장 예측 성능이 우수했던 **Gradient Boosting** 모델을 선정하여 Test Dataset 에 대해 **RMSE 0.084, R2 0.926** 을 기록하였다. 결과적으로 **최적 환경에서의 착과수는 내부 온도 관련 변수가 깊은 관련이 있다는 사실과 함께 착과 시기를 반영할 수 있는 시기 변수의 도입이 매우 성공적이었음을 확인하였다**.

본 팀은 나아가 사후 분석으로서, 시기 구분 선을 기준으로 나뉜 2~3 시기에 대해 **시기 별 회귀 모형 적합**을 실시하였다. 이를 통해 *각 시기 별 주요한 내부 변수를 파악하였으며, 상관 분석 및 도메인 지식을 토대로 하여 해당 내부 변수들과 외부 변수, 제어 변수 간의 관계를 설명* 하고자 하였다. 이를 통해 **각 시기에 주요하게 관리해야 할 요인이 무엇인지를 농업인에게 제시**할 수 있었다. 착과수와 기타 생육 변수에 대해서는 **착과수 자기회귀 모형 예측 및 타 생육 변수와 내부 변수 간 관계 분석**을 실시하였다.

이러한 일련의 과정에서 도출된 아이디어 및 한계점, 문제 의식을 구체화하여 **구조 방정식 모형, 최적화 모델 설계, 인과 추론 매커니즘 도입을 추후 연구 주제로 제안**하였다. 프로젝트를 진행하는 과정에서 본 팀은 **해당 분야에 대한 양질의, 다양한 데이터 축적이 그 무엇보다 중요하다는 필요성**을 느꼈고, 그 자세한 이유에 대해 제언 섹션에서 논의하며 보고서를 마무리하고자 한다.

2. 제언

본 팀이 도입한 '**시기 변수**'는 모델의 성공적인 착과수 예측에 중요한 영향을 미쳤다고 판단되나, 이러한 우수 성능이 미리 예상되고 계획되었다고 보기는 어렵다. 일단 시기 구분선이라는 개념 자체가 *시기에 따라 일정한 (최대 2~3 차) 함수의 형태로 착과수가 변화될 것이라는 강한 전제를 깔고 있다*. 가장 성능이 좋았던 최종 모델인 Gradient Boosting 모델은 학습 시 EDA 를 통한 단계 구분을 사용한 데이터를 사용하였는데, 이 방식은 *착과수와 어느정도로 관련이 있는지 모르는 두 개의 생육 변수 시각화 결과에 기반을 두고 있다*. 따라서 시기 구분은 이 잎 정보의 상호 상관적, 또는 상하적 변화를 설명할 수는 있어도 그것이 착과수 변화와 분명한 관계가 있음은 보장할 수 없다.

만약 가정 사항이 명백한 거짓이었고, 우리가 선택한 기준선이 착과수의 값 변동과는 전혀 관련이 없는 기준선이었다면 이와 같은 좋은 성능을 도출하지 못했을 것이다. 따라서, 시기 기준선 결정에 있어서 (Test 데이터를 제외한) **착과수 자체의 추세를 반영할 수 있는 방식을 사용할 필요**가 있으며, 이는 **착과수에 대한 데이터의 수를 늘리는 것** 만으로도 어느정도 달성될 수 있는 문제이다. **즉 6 개월에 대한 데이터가 아닌 5 개년에 대한 데이터가 확보**된다면, 데이터에서 나타나는 **착과수의 변화 추세 및 계절성을 판단하고 해당 년도의 외부 기상 데이터를 반영하여 착과량이 많을, 중간일, 적을 시기를 기준 년도에 대해서도 보다 정확하게 결정**할 수 있을 것이다. 이 외에도 딸기 재배에 대한 경험적 지식을 예측 모델에 어떻게 반영할 지에 대한 고민이 추가적으로 이루어져야만 한다.

본 팀이 아이디어로서 제안한 모델들도 구축을 위해서는 결국 **많은 양의 데이터가 필요**하다. 물론 본 팀이 수행한 것과 같이, 제한적인 데이터로도 해당 농가의 생육 상황을 분석하고 최적 환경 하에서의 착과수를 예측하는 모델링은 가능할 수 있다. 그러나 농업인들에게 보다 더 실질적인 도움을 주는 분석, 예컨데 *최적화 모형*과 *인과 추론 모형* 적합을 위해서는 여러 다른 상황에 대한 데이터 확보가 필수적이다. 비가 너무 많이 내려 제어 변수 처치로도 올바른 내부 환경 조성을 해주지 못했던 데이터, 특정 날짜에 난방 시스템이 잘못되어 온도가 지나치게 높아져 이후 착과수에 영향을 미친 데이터, 농사를 처음 시작해 경험이 부족하다보니 만족스러운 착과수를 확보하지 못했던 데이터 등이야 말로 모형의 목적에 부합한다. 이와 같은 **데이터의 양적 향상이 선행되어야 모델이 다양한 상황에서 원활하게 대처하여 농업인의 의사 결정을 도울 수 있고**, 보다 robust 한 착과수 예측을 내놓을 수 있을 것이다.

이전의 엽수, 줄기 등 생육 정보를 바탕으로 이후의 기준일 착과수를 예측하고 생장 상태를 감독하는 *자기 회귀적 모형*과 *구조 방정식 모형*은, 구축에 있어서 **동일 농가에서 동일 시기에 수집한 생육 및 환경 변수 정보**를 요구한다. 현재 스마트팜 빅데이터 플랫폼 상에는 생육 변수를 한 번에 수집한 이와 같은 형태의 데이터가 존재하지 않으며, '최적 환경'에 대한 기준 설명이 없고 어떤 기준으로 어떤 생육 변수를 동일 농가에서 획득했는지 파악하기 어렵다. 즉, 현 상태에서 **데이터의 질적 향상이 동반되어야만 해당 모델의 구현이 가능**하다.

따라서 데이터 플랫폼은 (1) **농업인들이 자신들의 데이터를 제공할 필요성을 직접적으로 체감**하고, (2) **젊은 연구자들이 해당 분야에 대한 새로운 아이디어를 활용하여 데이터를 다룰 수 있도록** *지속적으로 양쪽의 관심을 환기시켜주는 역할을 수행할 수 있어야 한다.* 이를 통해 양질의 데이터를 확보하여 다양한 형태의 분석을 허용해 주어야만 데이터 기반 스마트팜 산업은 기존의 패러다임에서 벗어난 발전을 이룩할 수 있을 것이다. 본 프로젝트가 이와 같은 '알깨기'에 조금이나마 기여할 수 있었기를 바라며 보고서를 마친다.

VI. 참고 문헌

저널

이종원. (2021). 데이터 기반의 스마트팜 연구 현황 및 발전방안. 한국농업기계학회 학술발표논문집. 26(2):25

유지혜, & 김태영. (2022). 인공지능 기반 스마트팜 작물 생육관리 의사결정 지원 모델 연구 현황 . 한국원예학회 학술발표요지.71-71.

천예원. (2023). *스마트팜 기술 현황과 표준화 동향 분석: 국내·외 비교분석을 중심으로 / Analysis of Smart Farm Technology Status and Standardization Trends*.

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

백진동. (2020). 빅데이터 기반의 딸기 생육환경 의사결정 시스템. *차세대융합기술학회논문지*, *4*(3), 258-264.

김나은, 한희선, 문병은, 최영우, & 김현태. (2022). 머신러닝 알고리즘을 이용한 온실 딸기 생산량 예측. *생물환경조절학회지*, *31*(1), 1-7.

홍성은, 박태주, 방준일, & 김화종. (2020). ConvLSTM 을 사용한 토마토 생산량및 성장량 예측 모델에 관한 연구. *한국정보기술학회논문지*, *18*(1), 1-10.

정호정, 노일래, & 김병수. (2013). 딸기'대왕','싼타','옥매','설향'및'매향'품종의 수경재배시 착과수 조절 효과. *경북대농학지*, *31*(4), 265-271.

이인하, 김현숙, 남명현, & 이병주. (2021). 신품종 딸기 '하이베리'품종의 착과수 조절이 생육 및 수량에 미치는 영향. *한국원예학회 학술발표요지*, 68-68.

김은완, 최정훈, 김보우, & 서동준. (2022). 스마트팜 이미지 데이터를 이용한 딸기의 생육 단계 분류 모델 개발에 관한 연구. *한국통신학회 학술대회논문집*, 1002-1003.

이세연, 양현정, 김민영, 김준경, 손아영, & 홍성훈. (2023). 스마트팜 활용을 위한 BI-LSTM 기반의 토마토 생산량 예측에 관한 연구. *한국통신학회논문지,* 48(4), 457-468.

이서희, 임종태, 차승우, 최지현, 최환용, 정상준, 황현중, 장준혁, RETITI DIOP EMANE Christopher, 김남영, 오영호, 김윤아, 유재수. (2022). *스마트 온실 환경 제어 의사결정 모델을 위한 딸기 생육 모델 분석*. 한국콘텐츠학회 종합학술대회 논문집, 개최지.

김문주, & 전민희. (2017). *Bayesian Structural Equation Modeling for Analysis of Climate Effect on Whole Crop Barley Yield*.

웹사이트

민선형, & 이준형. (2023). 디지털 농업을 위한 데이터 활용도 제고 방안 및 시사점. 한국농촌경제연구. https://repository.krei.re.kr/bitstream/2018.oak/29807/1/PRI099.pdf

디지털 농업을 위한 데이터 활용도 제고 방안 및 시사점. (2022, January 23). 경기도농수산진흥원 로고. https://gafi.or.kr/web/board/boardContentsView.do?contents_id=2f023609b7a04e7aafa617acc44d6dfe&board_id=60&menu_id=6e11cc2ff6114230b48717995b701546

https://github.com/Jithsaavvy/Explaining-deep-learning-models-for-detecting-anomalies-in-time-series-data-RnD-project/blob/master/notebooks/LIME_and_SHAP_Implementation.ipynb

https://www.kaggle.com/code/phamvanvung/shap-for-lstm/notebook

https://velog.io/@dlwns97/LSTM 시계열-데이터-예측