Regular paper | Journal of the KIECS. pp. 1027-1034, vol. 19, no. 5, Oct. 31. 2024, t. 127, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.5.1027

스마트팜 데이터 품질 향상을 위한 이상치 및 결측치 보정 방법에 관한 연구

이성재* · 심현

Research on Outlier and Missing Value Correction Methods to Improve Smart Farm Data Quality

Sung-Jae Lee* · Hyun Sim

요 약

본 연구는 AI 기반 스마트팜에서 발생하는 이상치 및 결측치 문제를 해결하여 데이터 품질을 향상시키고, 농업 예측 활동의 정확도를 높이는 것을 목표로 한다. 농진청·농정원에서 제공한 실제 데이터를 활용하여, 이상치 탐지 및 결측치 보정 기법을 적용함으로써 양질의 데이터를 수집하고 관리하고자 하였다. 성공적인 스마트팜 운영을 위해서는 IoT 기반의 AI 자동 생육 측정 모델이 필요하며, 이를 위해 안정적인 데이터 전처리를 통해 높은 데이터 품질 지수를 달성하는 것이 필수적이다. 본 연구에서는 생육 데이터의 이상치 및 결측치를 보정하는 다양한 방법을 적용하였으며, 제시된 데이터 전처리 방안을 머신러닝 기법을 통해 성능 평가 지수로 검증하였다. 연구 결과, 이상치 및 결측치 보정 방법을 적용한 결과 모델 성능이 크게 향상되었고, ROC와 AUC와 같은 평가 지표에서 높은 예측 정확도를 확인할 수 있었다.

ABSTRACT

This study aims to address the issues of outliers and missing values in AI-based smart farming to improve data quality and enhance the accuracy of agricultural predictive activities. By utilizing real data provided by the Rural Development Administration (RDA) and the Korea Agency of Education, Promotion, and Information Service in Food, Agriculture, Forestry, and Fisheries (EPIS), outlier detection and missing value imputation techniques were applied to collect and manage high-quality data. For successful smart farm operations, an IoT-based AI automatic growth measurement model is essential, and achieving a high data quality index through stable data preprocessing is crucial. In this study, various methods for correcting outliers and imputing missing values in growth data were applied, and the proposed preprocessing strategies were validated using machine learning performance evaluation indices. The results showed significant improvements in model performance, with high predictive accuracy observed in key evaluation metrics such as ROC and AUC.

키워드

Data Evaluation Model, Data Handling, Data Mining, Missing Value Correction, Outlier Removal
데이터 평가 모델, 데이터 처리, 데이터 마이닝, 누락값 수정, 이상치 제거

* 순천대학교 스마트농업전공(lsjtime@hanmail.net)
교신저자 : 국립순천대학교 스마트농업전공학과
· 접 수 일 : 2024. 08. 27
· 수정완료일 : 2024. 09. 19
· 게재확정일 : 2024. 10. 12
· Received : Aug. 27, 2024, Revised : Sep. 19, 2024, Accepted : Oct. 12, 2024
· Corresponding Author : Hyun Sim
Dept. Smart Agriculture, Sunchon National University
Email : simhyun@scnu.ac.kr

Ⅰ. 서 론

본 연구는 AI 기반의 스마트팜 데이터를 활용하여 생육 재배 예측 활동을 지원하고, 데이터 품질 기반의 관리를 통해 이상치 및 결측치를 보정함으로써 양질의 데이터를 수집하고 관리하는 것을 목표로 한다. 스마트팜의 성공적인 운영을 위해서는 IoT 기반 농가에서 수집되는 생육 데이터를 AI 기반 자동 측정 모델에 적용해야 하며, 이를 위해서는 데이터 핸들링 과정을 통해 높은 수준의 데이터 품질지수가 필수적이다. 스마트팜 데이터를 효과적으로 관리하기 위해서는 데이터의 일관성, 정확성, 그리고 완전성이 보장되어야 한다. 이를 위해, 데이터 수집 단계부터 체계적인 데이터 품질 관리 절차를 구축하고, 실시간으로 데이터를 모니터링하며 이상치 및 결측치를 자동으로 감지하고 보정하는 시스템이 필요하다. 또한, 수집된 데이터를 효율적으로 통합하고 분석함으로써 농업 생산성을 향상시킬 수 있는 방법을 지속적으로 모색할 필요가 있다. 본 연구에서는 다양한 생육 데이터에서 발생하는 이상치 및 결측치 보정 방안을 구체적으로 적용하였다. 사례별로 사용하기 쉽게 보정 방안을 제공하고, 이 방안에 따라 수집된 데이터를 머신러닝 기법을 통해 분석하였다. 머신러닝 모델의 성능은 다양한 평가 지수를 사용하여 검증되었으며, 모델 성능을 ROC와 AUC 측정치를 통해 평가하였다. 이를 통해 보정된 생육 데이터의 신뢰성과 예측 정확도를 검증하였다. 연구 결과, 생육 데이터의 이상치 및 결측치 보정 방법을 적용한 경우, 모델 성능 평가 지수, ROC 곡선, 및 AUC 값에서 모두 높은 예측 정확도를 확인할 수 있었다. 특히, 실시간 데이터 마이닝 결과에서도 보정된 데이터가 효과적으로 활용되었음을 확인하였다. 이러한 연구 결과는 스마트팜 데이터 관리와 분석 과정에서 데이터 품질 관리가 중요함을 강조하며, 고품질 데이터를 기반으로 스마트팜 운영의 효율성을 높이는 데 기여할 수 있음을 시사한다. 본 연구를 통해 스마트팜 데이터의 품질 향상을 위한 체계적인 관리 방안을 제시하였으며, 이를 통해 스마트 농업의 효율성과 생산성을 크게 향상시킬 수 있는 가능성을 확인하였다. 결론적으로, AI 기반의 데이터 관리 및 분석 기법을 활용하여 생육 예측 정확도를 높이고, 스마트팜 운영의 성공적인 모델을 구축하는 데 있어 본 연구의 기여가 클 것으로 기대된다.

Ⅱ. 관련 연구

2.1 스마트팜 데이터 품질 향상: 이상치 제거

스마트팜 데이터의 품질을 향상시키기 위해 다양한 이상치 탐지 기법이 활용될 수 있다. 대표적인 이상치 탐지 기법과 이를 스마트팜 데이터에 적용한 연구 사례는 다음과 같다.

먼저, 표준 점수(Z-Score)는 데이터 포인트가 평균에서 몇 표준편차만큼 떨어져 있는지를 측정하는 방식이다. 일반적으로 Z-Score가 ±3을 초과하는 경우를 이상치로 간주한다[1]. Z-Score는 스마트팜 환경 데이터(예: 온도, 습도)에서 극단값을 제거하는 데 효과적으로 사용되며, 데이터의 정규성을 확보하여 분석의 정확도를 향상시키는 장점이 있다.

사분위 범위(IQR: InterQuartile Range)는 데이터의 1사분위(Q1)와 3사분위(Q3) 사이의 범위를 이용하여 이상치를 판별하는 기법이다[2]. 일반적으로 IQR의 1.5배를 벗어나는 값을 이상치로 간주하며, 이는 작물 생육 데이터에서 비정상적으로 높거나 낮은 값을 탐지 및 제거하는 데 유용하다. 이 기법은 데이터의 분포를 고려한 유연한 이상치 탐지가 가능하다는 점에서 그 장점이 있다.

상자그림(Boxplot)은 IQR을 기반으로 한 시각적 이상치 탐지 기법으로, 상자 외부에 위치한 데이터 포인트들을 이상치로 간주한다. 이 방법은 토양 수분 함량, 영양분 농도 등 다양한 스마트팜 데이터에서 이상치를 탐지하는 데 활용되며, 직관적인 시각화를 통해 이상치를 쉽게 확인할 수 있다는 장점이 있다.

밀도 기반 클러스터링(DBSCAN: Density-Based Spatial Clustering of Applications with Noise)은 밀도 기반 클러스터링 기법을 활용하여 밀도가 낮은 영역의 데이터 포인트를 이상치로 간주하는 방식이다[3]. 이 기법은 스마트팜 센서 데이터에서 노이즈나 오류 데이터를 제거하는 데 효과적이며, 복잡한 형태의 클러스터에서도 이상치를 탐지할 수 있다는 장점을 지닌다.

이러한 이상치 탐지 기법들은 스마트팜 데이터의 특성 및 분석 목적에 맞추어 적절히 선택하여 적용할 수 있다. 이를 통해 스마트팜 데이터의 품질을 향상시킴으로써 AI 모델의 성능을 높이고, 스마트팜 데이터 전처리의 효율성을 제고할 수 있다. 또한, 각 기법은

장단점이 있으므로, 필요에 따라 여러 기법을 조합하여 사용하는 것이 유용할 수 있다.

2.2 스마트팜 데이터 품질 향상: 결측치 제거

스마트팜 코드 및 데이터 통합체계 구축 과정에서, 농촌진흥청 및 농정원의 품질기준을 통합하고, 수집된 데이터의 활용 가치를 극대화하기 위해 결측치 처리 및 이상치 제거를 통한 품질관리가 필수적이다. 특히, 환경 및 제어 데이터를 대상으로, 일일 30% 이내의 결측이 발생한 농가 데이터는 결측치를 보정하여 분석 서비스 등에 활용할 수 있다. 이를 위해, 결측 시점의 1시간 전 측정 데이터와 동일 농가 내 1~5일 이전 동일 시간대의 1시간 간격 변화량의 평균값을 합산하여 결측치를 보정하는 기준을 마련하였다. 이러한 결측치 처리 기준은 데이터 품질 평가 지표로도 활용되며, 스마트팜 데이터 전처리에서 중요한 요소로 작용한다.

2.3 결측치 처리 방법

스마트팜 데이터의 전처리 과정에서 결측치 처리는 매우 중요한 단계이며, 다양한 처리 방법이 존재한다. 첫째, **결측치 삭제 방법**은 가장 기본적인 처리 방법으로, 결측치가 포함된 행이나 열을 삭제하는 방식이다. 행 삭제는 결측치가 포함된 레코드(행)를 제거하며, 열 삭제는 결측치가 다수 존재하는 변수를 제거한다. 이 방법은 적용이 간단하지만, 데이터의 크기가 줄어들어 분석 신뢰성이 저하될 가능성이 있다. 따라서 결측치가 무작위로 발생하고, 그 양이 매우 적을 때 적합한 방법이다.

둘째, **통계적 대체 방법**은 결측치를 삭제하지 않고 통계량을 활용해 대체하는 방법이다. **평균 대체**는 결측 메커니즘이 완전 임의 결측(MCAR)일 때 유용하며, 결측치를 해당 변수의 평균값으로 대체한다. 반면, **중앙값 대체**는 결측치를 해당 변수의 중앙값으로 대체하는 방법으로, 데이터가 비대칭 분포를 가질 때 유용하다. 이 방법은 많은 프로그램에서 기본적으로 제공되지만, 분석 결과에 왜곡을 초래할 수 있다.

셋째, **회귀 분석을 통한 대체 방법**은 회귀 모형을 이용하여 결측치를 추정하는 방식이다. 먼저 결측치를 임시로 통계량(예: 평균값)으로 대체한 후, 결측치가 있는 변수를 목표변수로 설정하고 나머지 변수를 설명변수로 하여 회귀모형을 구축한다. 이후 이 회귀모형을 사용하여 결측치를 예측하고 대체하며, 변수 간 강한 상관관계가 있을 때 효과적이다. 이 과정은 반복적으로 수행되어 결측치를 지속적으로 최신화할 수 있다[4].

넷째, **머신러닝 기반 대체 방법**은 머신러닝 알고리즘을 이용하여 결측치를 예측하고 대체하는 방법이다. 먼저 결측치를 통계량으로 임시 대체한 후, 결측치가 있는 변수를 목표변수로 설정하고 나머지 변수를 설명변수로 하여 머신러닝 모델을 구축한다. 이 모델을 통해 결측치를 예측하여 대체하며, 대규모 데이터셋 및 복잡한 관계를 처리하는 데 유용하다. 모델의 설명력(예: R^2)이 일정 기준을 충족하면 최종 예측값으로 결측치를 대체하게 된다.

다섯째, **최빈값 대체법**은 범주형 변수의 결측치를 해당 변수의 최빈값으로 대체하는 방식이다. 이 방법은 범주형 변수에 적합하며, 연속형 변수에는 적합하지 않다. 따라서 주로 범주형 데이터에서 결측치 처리를 위해 사용된다.

여섯째, **다중 대체법(Multiple Imputation)**은 결측치를 여러 번 대체하여 불확실성을 고려하는 방식이다. 다중 대체법은 여러 가능한 값들로 결측치를 여러 번 대체하며, 결측 메커니즘이 임의 결측(MAR)일 때 적합하다. 이 방법은 편향을 줄일 수 있는 장점이 있지만, 계산 비용이 높고 복잡하다.

마지막으로, **예측 모델 기반 대체(Model-based Imputation)**는 머신러닝 모델을 사용하여 결측치를 예측하고 대체하는 방식이다. 이 방법은 복잡한 패턴을 포착하는 데 유리하며, 대규모 데이터셋과 복잡한 변수 간 관계를 처리할 때 유용하다. 그러나 모델 선택과 튜닝 과정이 필요하고, 계산 비용이 높다는 단점이 있다. 이처럼 스마트팜 데이터 분석에서 결측치 처리 방법은 다양한 방식으로 적용될 수 있으며, 데이터의 특성에 맞는 적절한 결측치 처리 기법을 선택하는 것이 데이터 품질과 분석의 신뢰성을 보장하는 중요한 요소이다.

2.4 스마트팜 데이터 품질향상을 위한 품질지수

스마트팜 생육 데이터의 경우, 데이터의 품질지수를 활용한 완전성, 유일성, 유효성, 일관성, 정확성, 무결성의 6가지 지표[5]로 전처리를 한다. 실제 농가 데

이터에서 발생하는 데이터는 의미 없는 값이거나, 누락 및 오타가 발생하여 품질이 좋지 못하다. 품질이 낮은 데이터를 분석에 이용하면 좋은 결과를 얻기 힘들다. 그러므로 데이터 (품질) 전처리는 데이터의 분석에 있어서 필수적인 단계이다. 데이터 품질 전처리를 위해 아래 6가지의 데이터 품질지수를 파악하고, 데이터의 품질지수를 향상 시킨다.

• 완전성 품질지수= $(1-\frac{\text{결측데이터개수}}{\text{전체데이터개수}})*100$

완전성 항목의 스마트팜 데이터의 경우, 결측 데이터의 값이 30% 이상인 데이터들은 데이터의 완전성을 떨어뜨리기 때문 에 각 열(column)에 대한 결측 데이터값의 비율을 확인하여 삭제한다.[]

• 유일성 품질지수= $(\frac{\text{유일한 데이터개수}}{\text{전체 데이터개수}})*100$

데이터에서 유일한 값을 갖는 열(column)을 찾는다.

• 유효성 품질지수= $(\frac{\text{유효성을 만족한 데이터의 개수}}{\text{전체 데이터의 개수}})*100$

'max()', 'min()' 함수를 이용하여 데이터가 유효범위를 벗어나지 않는지를 확인한다.

• 일관성 품질지수= $(\frac{\text{일관성을 만족한 데이터의 개수}}{\text{전체 데이터의 개수}})*100$

일관성 데이터 테이블 간 종속관계를 확인한다.

• 정확성 품질지수

$= 1-(\frac{\text{정확성에 위배되는 데이터의 개수}}{\text{전체데이터의 개수}})*100$

데이터의 '평균값'이 수집된 데이터에 한하여, 다음과 같이 정확성 품질지수를 확인 할 수 있다.

• 무결성 품질지수

$=\left\{1-(\frac{\text{유일성,유효성,일관성중 100\%가아닌지수}}{3})\right\}*100$

무결성 품질지수는 데이터가 '유일성', '유효성', '일관성' 지수를 만족하는지 확인하는 지표이다.

일반적으로 세 가지 지수 중 100% 품질을 담당하는 지수의 비율을 무결성 품질지수로 정의한다.

III. 실 험

3.1 연구 방법 및 성능 평가

앞선 선행연구에서는 각 알고리즘마다 임의로 생성한 불완전 데이터를 이용하여 대체값을 생성하고 참값과 비교하여 성능을 평가하였다. 이 과정에서 연속형 변수와 범주형 변수를 함께 측정할 수 없는 한계가 있었으며, 이에 따라 다른 검증 방법을 사용하여 성능을 비교하였다[6][1]. 연속형 변수의 결측치 대체 성능은 주로 RMSE(: Root Mean Squared Error)를 표준화하여 검증하였으며, 범주형 변수의 성능 평가는 혼동 행렬 (Confusion Matrix) 기반 Accuracy(정확도)를 사용하여 모델 성능 평가 지표로 활용하였다. 또한, 각 알고리즘이 이상치 및 결측치를 처리하는 데 걸리는 시간을 기준으로 성능 평가를 비교하였다.

본 연구에서는 결측치와 이상치가 포함된 불완전한 스마트팜 실증 데이터를 대상으로, 전처리 전후의 데이터 분석 결과를 비교하였다. 결측치 및 이상치를 포함한 원본 데이터와 전처리 후 정제된 데이터를 분석하여, 각각의 성능 평가를 수행하였다. 수치 예측의 경우 RMSE 및 R-squared 지표를 사용하여 성능을 검증하였으며[7], 이를 통해 스마트팜 데이터의 품질지수가 전처리 전후로 얼마나 차이가 있는지를 평가하였다. 표 1에서는 토마토 생육 데이터의 일반 현황을 나타낸다. 토마토 생육 데이터를 활용하여 이상치 처리 및 결측치 보정 전후의 데이터 성능을 평가한 결과, 전처리가 모델 성능에 미치는 영향을 확인할 수 있었다.

표 1. 토마토 생장 데이터의 일반 상태
Table 1. General status of tomato growth data

No	Category	Contents	Notes
1	Collection type	csv	
2	Number of data	17*931=15,827	Numerical
3	Missing values	5%(Number of fruits)	Numerical

전처리 전후의 데이터 성능평가 프로세스로 아래 표2의 단계별 진행하였다.

1) http://www.riss.kr/link?id=T15480243&outLink=K83.

표 2 토마토를 이용한 전처리(이상치, 누락값 수정) 전/후 데이터 비교 과정
Table 2. Comparison process of data before and after preprocessing (outlier, missing value correction) using tomato data

No	Contents
1	Load the necessary libraries
2	Read data and collect data before/after preprocessing
3	Process missing values and match the number of rows in the data
4	Train and evaluate multiple models (Random Forest, XGBoost, Lasso, SVM)
5	Compare performance before and after preprocessing
6	Output the results and calculate the performance improvement
7	Output the variable importance of the Random Forest model
8	Compare performance before and after preprocessing, compare R-squared, and visualize the performance improvement rate as a bar graph

그림 1에서는 각 모델(Random Forest, XGBoost, Lasso, SVM)의 전처리 전후 RMSE 및 R-squared 값을 테이블 형태로 출력하였다. RMSE_Improvement는 전처리 전후의 RMSE 개선 정도를 백분율로 계산한 값이며, 양수는 성능 개선을, 음수는 성능 저하를 의미한다[8]. 마찬가지로, R2_Improvement는 전처리 전후의 R-squared 개선 정도를 백분율로 계산하며, 양수는 성능 개선을, 음수는 성능 저하를 나타낸다. 각 모델별로 RMSE와 R-squared의 개선 정도를 백분율로 보여주며, 전처리 전후 모델 성능을 비교할 수 있다.

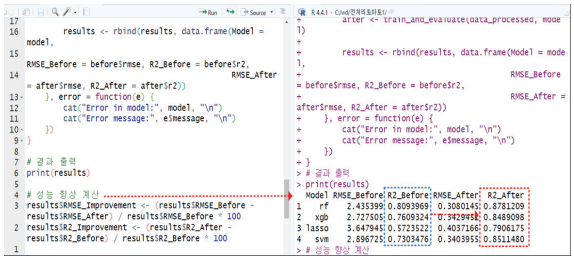

그림 1. 아키텍처 전처리 전/후 성능 개선 결과 화면
Fig. 1 ArchitecturePerformance improvement result screen before and after preprocessing

이 결과 해석 시 RMSE는 낮을수록, R-squared는 높을수록 좋은 성능을 의미한다. 특히, Random Forest 모델에서 각 변수의 중요도를 측정할 때, IncNodePurity는 각 변수가 노드 순도(node purity)를 얼마나 증가시키는지를 나타내며, 값이 클수록 해당 변수가 모델에서 더 중요한 역할을 한다는 것을 의미한다. 표 3은 전처리 후 모델 성능이 5% 개선된 항목을 나타내며, 성능 개선은 최대 38.13%에서 최소 8.5%까지의 개선 결과를 보였다. 이는 스마트팜 데이터에서 결측치 및 이상치 처리 후 성능 향상에 대한 구체적인 수치를 보여주는 중요한 결과로, 데이터 전처리의 필요성을 강조한다.

표 3 전처리 전/후 성능 개선 차트
Table 3. Performance improvement chart before and after preprocessing

Category	before RMSE	before R^2	after RMSE	after R^2	Improvement rate%
rf	2.44	0.81	0.31	0.88	8.5
xgb	2.73	0.76	0.34	0.85	11.56
lasso	3.65	0.57	0.40	0.80	38.13
svm	2.89	0.73	0.34	0.85	16.54

그림 2는 각각의 알고리즘의 RMSE수치를 전처리 전후로 나타내는 그림으로 표3의 결과를 시각화하여 막대그래프로 표현하였다.

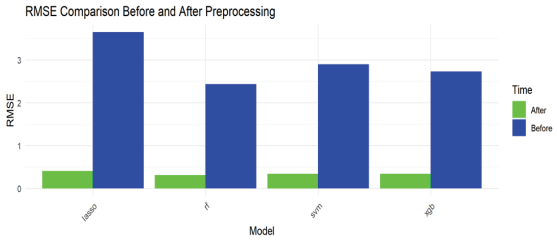

그림 2. RMSE 수치 전처리 전 후 비교표
Fig. 2 Comparison table before and after RMSE numerical preprocessing

아래의 그림 3은 전처리후의 향상율을 시각화하여 전후를 비교하여 시각적으로 전체 향상도를 표현한다.

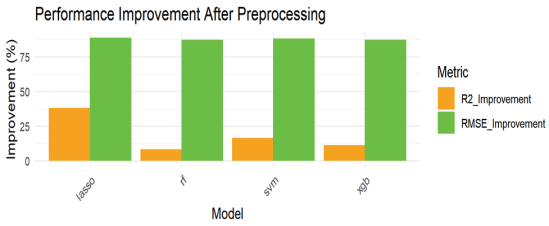

그림 3. 전처리 후 성능 개선
Fig. 3 Performance Improvement After Preprocessing

3.2 전처리 방법과 모델 성능 개선

관련 연구를 참고하여, 데이터셋의 특성에 따라 다양한 전처리 방법을 시도할 수 있다. 표준화(Standardization)와 정규화(Normalization)를 적용하여 데이터 전처리 과정에서 이상치를 조절하고, 평가 모델 구축 시 하이퍼파라미터 튜닝을 통해 각 모델의 성능을 개선할 방안을 고려하였다[9]. 이를 위해 그리드 서치(Grid Search)나 랜덤 서치(Random Search)를 사용하여 최적의 하이퍼파라미터를 탐색하고, 교차검증(Cross-validation)을 통해 모델 성능을 보다 안정적으로 평가할 수 있다. 이와 함께, 특성 중요도(Feature Importance)를 분석하여 예측에 가장 큰 영향을 미치는 변수를 파악하는 것이 가능하다.

그림 4와 그림 5는 앙상블 기법(Ensemble Methods)을 사용하여 여러 모델의 예측 결과를 결합함으로써 더 나은 성능과 예측 향상을 도출한 사례를 보여준다. 앙상블 기법을 적용한 결과, 93%의 예측 정확도를 얻을 수 있었다.

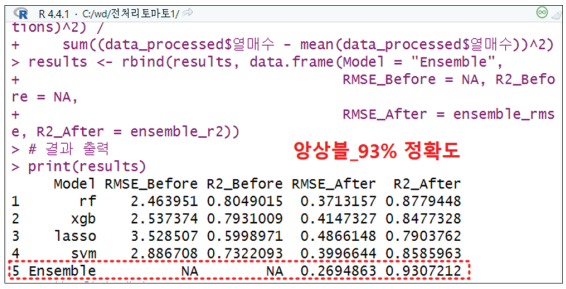

그림 4. 최종 피팅 알고리즘 정확도
Fig. 4 Final fitting algorithm accuracy

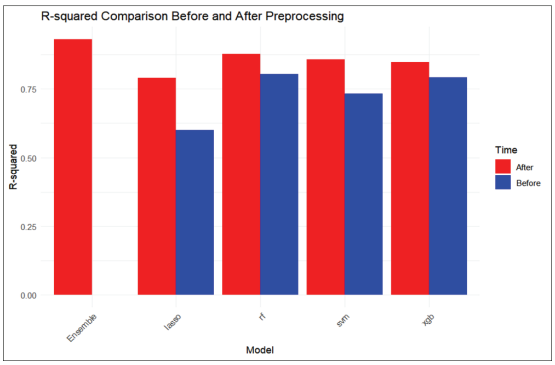

그림 5. 앙상블 모델 추가 성능 평가 그래프
Fig. 5 Ensemble model additional performance evaluation graph

3.3 데이터 품질지수와 예측 모델 성능의 관계

본 연구의 마지막 결과에서, 전처리 전후의 데이터 품질지수와 예측 모델 성능 간의 관계를 분석한 결과, 일부 품질지수는 전처리 후 향상된 것으로 나타났으나, 유일성과 무결성 등의 품질지수는 오히려 전처리 후에 낮아졌다. 이러한 결과를 바탕으로, "데이터 품질지수가 높다고 해서 반드시 예측 모형의 정확도가 높아지는 것은 아니다"라는 결론을 도출할 수 있었다[10]. 그 이유는 다음과 같이 추론할 수 있다.

1) 관련성 부족: 높은 품질의 데이터라도, 예측하려는 목표 변수와의 관련성이 낮을 수 있다. 데이터가 깨끗하고 완전하더라도, 예측에 유용한 정보를 포함하지 않을 수 있다.

2) 과적합의 위험: 지나치게 깨끗한 데이터는 과적합(overfitting)을 유발할 수 있다. 이는 실제 세계의 노이즈와 변동성을 반영하지 못하게 되어 모델의 성능 저하로 이어질 수 있다.

3) 복잡성 반영 부족: 데이터 품질지수는 데이터 간의 복잡한 상호작용이나 비선형 관계를 고려하지 않는다. 개별 특성들이 높은 품질을 유지하더라도, 이들이 조합되었을 때의 예측력을 보장하지 못한다.

4) 시간적 요소의 미반영: 시계열 데이터를 다룰 때, 단순한 품질지수는 시간에 따른 패턴이나 추세를 반영하지 못할 수 있다.

5) 외부 요인의 영향: 예측 모델의 성능은 사용된 알고리즘, 하이퍼파라미터 튜닝, 특성 선택 등 다양한 외부 요인에 의해 영향을 받을 수 있다.

6) 분포의 불균형: 데이터의 분포가 불균형하거나 편향되어 있는 경우, 품질지수가 높더라도 예측 성능이 저하될 수 있다. 이와 같이, 데이터 품질지수는 데이터의 특정 측면을 평가하는 데 유용하지만, 예측 모델의 성능을 보장하는 유일한 척도가 될 수 없음을 본 연구는 시사하고 있다. 데이터 전처리와 품질 관리뿐만 아니라, 모델 구축 과정에서의 종합적인 접근이 필요하다..

Ⅳ. 결 론

본 연구는 데이터 품질지수가 예측 모형 개발의 중요한 시작점이지만, 그 자체로 높은 예측 정확도를 보장하지 않는다는 결론을 제시하였다. 효과적인 예측 모델링을 위해서는 데이터 품질 향상뿐만 아니라, 적절한 특성 공학, 알고리즘 선택, 모델 튜닝, 그리고 도메인 지식의 활용이 종합적으로 이루어져야 한다.

전처리 과정을 통해 데이터의 전반적인 품질이 향상되었다. 결측치 처리, 이상치 제거, 특성 스케일링 등의 작업을 통해 데이터의 완전성, 일관성, 정확성이 개선되었으며, 특히 결측치 비율이 크게 감소하여 데이터의 완전성이 향상되었다. 또한, 스케일링을 통해 변수 간 척도 차이가 줄어들어 모델 학습에 긍정적인 영향을 미쳤다.

모델 성능 측면에서는 대부분의 알고리즘에서 성능 개선이 관찰되었다. Random Forest, XGBoost, SVM 모델에서 RMSE가 감소하고 R-squared 값이 증가하여 예측 정확도가 향상되었으며, 특히 Random Forest 모델의 성능 향상이 두드러졌다. 이는 전처리 과정이 트리 기반 모델에 특히 효과적이었음을 시사한다. 반면, Lasso 회귀 모델의 경우 성능 개선이 미미하였는데, 이는 Lasso가 이미 특성 선택 기능을 내재하고 있어 전처리의 영향을 덜 받았기 때문으로 추정된다. 앙상블 모델은 가장 높은 성능을 보여, 다양한 모델의 장점을 결합하는 접근이 효과적임을 확인하였다.

변수 중요도 분석 결과, '초장', '엽수', '줄기 굵기' 등의 변수가 '열매수' 예측에 중요한 역할을 하는 것으로 나타났다. 이는 농업 도메인 지식과 일치하는 결과로, 모델의 신뢰성을 뒷받침하는 중요한 발견이다.

종합적으로, 본 연구에서 수행된 전처리 과정은 데이터 품질과 모델 성능 향상에 긍정적인 영향을 미쳤다. 그러나 일부 모델에서는 성능 개선 효과가 제한적이었음을 감안할 때, 전처리 방법의 선택과 적용에 있어 모델 특성과 데이터 특성을 종합적으로 고려해야 한다는 점을 알 수 있다. 향후 연구에서는 더 다양한 특성 공학 기법과 고급 모델링 접근법을 탐색하여 예측 성능을 더욱 개선할 수 있을 것으로 기대된다.

감사의 글

본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화 혁신인재양성사업임(IITP-2024-RS-2020-II201489).

References

[1] A. B. B. Torres, A. R. da Rocha, T. L. Coelho da Silva, J. N. de Souza, and R. S. Gondim, "Multilevel data fusion for the internet of things in smart agriculture," *Computers and Electronics in Agriculture,* vol. 171, 2020, pp. 105309.

[2] S. Mandić-Rajčević and C. Colosio, "Methods for the Identification of Outliers and Their Influence on Exposure Assessment in Agricultural Pesticide Applicators: A Proposed Approach and Validation Using Biological Monitoring," *Toxics,* vol. 7, no. 3, 2019, pp. 37. DOI: 10.3390/toxics7030037.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD-96),* 1996. DOI: 10.5555/3001460.3001507.

[4] S. Kim, "Efficiency of Imputation Methods in Generalized Estimating Equations," *Agriculture,* vol. 7, no. 3, , Aug. 2018, pp. 56-59.

[5] K.-R. Shon, "Research of Quality Improvement by Factors Analysis Data Quality Problem: Focus on National R&D Information Linking Structure," *Journal of the Korea Contents Association,* vol. 9, no. 1, 2009, pp. 23-28.

[6] S. Lee, "Comparison of Algorithms for the Missing Data Imputation Methods," *Report,* 2020, pp. 44-45.

[7] Kakao, "Kakao AI Report," *Report,* Oct. 2017, pp. 1-15.
[8] J. Yu, K. Jung, Y. Chung, and C. Lee, "A Study on the Prediction Model of the Radius of Curvature of the Subtle Feature of the Automotive Parts for Different Forming Conditions," *J. Korean Soc. Precis. Eng.,* vol. 40, no. 1, 2023, pp. 49-55.
DOI: 10.7736/JKSPE.022.101.
[9] W. Chung, O. Moon, S. Park, and E. Hwang, "An Electrical Load Forecasting Model based on GNN Considering Spectral Similarity and Priori Relationship," in *Proc.* Korean DataBase Conference, 2022, pp. 3-6.
https://www.dbsociety.kr/kdbc/kdbc2022/KDBC2022_Proceedings.pdf.
[10] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Scientific Reports,* vol. 14, Article no. 6795, 2024.
DOI: 10.1038/s41598-024-54807-1.

저자 소개

이성재(Sung-Jae Lee)

2023년~현재 순천대학교 스마트농업전공 재학

1999년~현재 (주)태광에너지 근무

※ 관심분야 : 스마트팜, 인공지능

심현(Hyun Sim)

2002년 순천대학교 컴퓨터과학과 졸업(이학석사)
2009년 순천대학교 대학원 컴퓨터과학과 졸업(이학박사)
2020년~현재 순천대학교 스마트농업전공 부교수
2023년~현재 순천대학교 정보전산원장
2021년~현재 순천대학교 산학협력교육 센터장
2021년~현재 HyperAi&ESG검증심사연구소 소장
※ 관심분야 : 인공지능, 스마트팜, ESG