Journal of Research in Mechanics of Agricultural Machinery

Shahrekord University

Research Paper

Sugarcane Yield Prediction through Integrated Satellite Sentinel-2 Indices and Management Features using Ensemble Machine Learning Algorithms

Feryal Jaderi [1], Nasim Monjezi [1*]

[1.] Department of Biosystems Engineering, Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran.

Article History

Submited:2026/04/20

Revised: 2026/06/18

Accepted: 2026/07/16

Published online: 2026/09/23

Keywords:

Crop yield prediction
Remote sensing
K-means clustering
Ensemble machine learning
Precision agriculture

***Corresponding author email:**
n.monjezi@scu.ac.ir

ORCID:
0000-0001-8229-7706

Abstract

Accurate prediction of sugarcane yield plays a crucial role in improving farm management and supply chain sustainability. In this research, an accurate framework for predicting sugarcane yield in the Dakht-e-Dehkhoda Sugarcane Agro-Industry (Khuzestan Province) has been developed, based on the combination of satellite data and agricultural information. The data used included 2417 records from the agricultural years 1396 to 1403, gathered from farm management records and Sentinel-2 satellite imagery, including standard vegetation indices such as NDVI and EVI. In the first step, farms were categorized into four distinct groups using the K-means clustering algorithm, based on their management and yield characteristics. Subsequently, to enhance model accuracy, engineered features like water use efficiency and fertilizer use efficiency were defined. Following this, two machine learning models of the Random Forest and Gradient Boosting types were trained for yield prediction. Model evaluation was performed by allocating 80% of the data for training and the remaining 20% to an independent test set. To ensure model stability and generalizability, 5-fold cross-validation was employed during the training phase. The results indicated that the Gradient Boosting model achieved the best performance, reaching a coefficient of determination of 0.9924 and a root mean square error of 1.88 tons per hectare. Furthermore, feature importance analysis revealed that water use efficiency, with a share of 87%, was the most effective factor in yield prediction. In conclusion, the findings of this research confirm that combining management data with satellite information and employing feature engineering can serve as a precise tool to support spatial decision-making and resource management in precision agriculture.

How to cite this paper:

Jaderi, F. and Monjezi, N. (2026). Sugarcane Yield Prediction through Integrated Satellite Sentinel-2 Indices and Management Features using Ensemble Machine Learning Algorithms. *Journal of Research in Mechanics of Agricultural Machinery.* 40: 31-48. https://dx.doi.org/10.22034/jrmam.2026.14989.762 (In Persian)

https://dx.doi.org/10.22034/jrmam.2026.14989.762

EXTENDED ABSTRACT

Introduction

In Iran, Khuzestan Province is the primary hub for sugarcane production, holding significant importance for national food security and the sugar industry. Due to the vast scale of production in this region, even minor improvements in yield prediction accuracy can lead to substantial economic savings and more efficient resource management at both regional and national levels. In recent decades, precision agriculture technologies, especially remote sensing and machine learning, have offered promising solutions to these challenges. Satellite imagery, such as the high-resolution data from the Sentinel-2 mission, enables continuous and large-scale monitoring of crop health and growth. The main objective of this research is to develop and evaluate a high-accuracy yield prediction framework for sugarcane by integrating satellite and agro-technical data. The key innovation of this study lies in two aspects: first, the introduction and validation of novel engineered features, such as "Water Use Efficiency," that directly model the impact of farm management. Second, a comparative analysis of two powerful ensemble models, Random Forest and Gradient Boosting, is conducted to identify the best approach for achieving maximum predictive accuracy.

Material and Methods

This research was conducted on sugarcane farms of the Dekhoda Sugarcane Agro-industry in Khuzestan Province, Iran. The dataset integrates three distinct data sources: Agro-technical and Management Data: This dataset includes 2,417 records of farm plots for the crop years 2017 to 2024, extracted directly from farm records. Key variables include final gross yield (tons/ha), water consumption (m^3/ha), total fertilizer application (kg/ha), and soil electrical conductivity (EC). Satellite Data: We used Sentinel-2 Level-2A (atmospherically corrected) multi-spectral images. Meteorological Data: Daily climatic variables including average temperature, relative humidity, wind speed, precipitation, and sunshine hours were obtained from the nearest weather station and averaged for the growing season. Three standard VIs were calculated from Sentinel-2 images: Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). To identify spatial patterns and group plots with similar characteristics, the K-means clustering algorithm was used. Two powerful ensemble models, Random Forest and Gradient Boosting, were trained to predict final yield. All data analysis and modeling were performed using Python (version 3.13.2) and its key libraries, including Pandas and Scikit-learn. The dataset was split into 80% training and 20% test sets using a stratified approach based on the cluster labels. Model stability was evaluated using 5-fold cross-validation. Performance was measured using R^2, RMSE, and MAE.

Results and Discussion

The statistical analysis of the final dataset, comprising 2,417 sugarcane plots, revealed significant yield variability, with a standard deviation of 22.17 tons/ha, highlighting the inherent complexity of the prediction task. The K-means clustering successfully identified four distinct agro-technical profiles. For instance, Cluster 1, with the highest average yield (76.78 tons/ha), low soil salinity, and highest fertilizer use, represented high-potential plots with intensive management. Comparison of Prediction Models: The performance of both Random Forest and Gradient Boosting models was evaluated on the unseen test data. The results, summarized in Table 3, show that both models achieved high accuracy, but the Gradient Boosting model exhibited superior performance, with an R^2 of 0.9924 and an RMSE of 1.88 tons/ha. This performance is on par with the state-of-the-art in this field. The high accuracy of the Gradient Boosting model is largely attributable to the feature engineering step. A feature importance analysis revealed that the water efficiency feature was the most dominant predictor,

accounting for 87% of the model's explanatory power. This finding is in contrast to studies that focus solely on radar data and underscores the critical importance of integrating on-ground management data with optical satellite data. Performance Analysis within Clusters and Comparison with Prior Work: The model's performance was also evaluated within each cluster to assess its stability. The results, shown in Table 5, demonstrate that the Gradient Boosting model maintained consistently high accuracy (R^2>0.96) in the two largest clusters (0 and 1). This indicates that the model is robust and performs well under varying conditions, such as high yield or high salinity

Conclusions

This research presented a comprehensive framework for high-accuracy sugarcane yield prediction by integrating satellite and agro-technical data. The study showed that by using advanced feature engineering and ensemble machine learning models, highly accurate yield estimations can be achieved. Farm plots were first clustered using K-means, and then two ensemble models, Random Forest and Gradient Boosting, were trained on a rich set of features. The main achievement of this research was reaching a state-of-the-art performance level with the Gradient Boosting model, which achieved an R^2 of 0.9924. The key innovation and the reason for this success was the introduction and validation of engineered features. The feature importance analysis revealed that the "water efficiency" feature alone accounted for 87% of the model's explanatory power, emphasizing the critical role of efficient water management compared to standard vegetation indices.

Acknowledgements

The authors would like to thank Shahid Chamran University of Ahvaz for providing funding for this research.

Author Contributions

Feryal Jaderi: Data collection and analysis, and initial writing

Nasim Monjezi: Study design, text revision, Providing expert opinions and reviews

Data Availability Statement

Not applicable

Ethical Considerations

The authors avoided data fabrication, falsification, plagiarism, and misconduct.

Conflict of Interest

The author declares no conflict of interest.

Funding Statement

The authors received no specific funding for this research

نشریه پژوهش‌های مکانیک ماشین‌های کشاورزی
دانشگاه شهرکرد

مقاله پژوهشی

پیش‌بینی عملکرد نیشکر با رویکرد یکپارچه‌سازی شاخص‌های ماهواره‌ای Sentinel-2 و ویژگی‌های مدیریتی، با استفاده از الگوریتم‌های یادگیری ماشین گروهی

فریال جادری[۱] و نسیم منجزی[۱]*

۱. گروه مهندسی بیوسیستم، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، اهواز، ایران.

تاریخچه مقاله

دریافت: ۱۴۰۵/۰۱/۳۱
بازنگری: ۱۴۰۵/۰۳/۲۸
پذیرش: ۱۴۰۵/۰۴/ ۲۵
انتشار: ۱۴۰۵/۰۷/۰۱

واژه‌های کلیدی:
پیش‌بینی عملکرد محصول سنجش‌ازدور
خوشه بندی K میانگین
یادگیری ماشین گروهی
کشاورزی دقیق

*پست الکترونیکی نویسنده مسئول:
n.monjezi@scu.ac.ir

ORCID: ۰۰۰۰-۰۰۰۱-۸۲۲۹-۷۷۰۶

چکیده

پیش‌بینی دقیق عملکرد نیشکر نقشی اساسی در بهبود مدیریت مزارع و ارتقای پایداری زنجیره تأمین دارد. در این پژوهش، چارچوبی برای پیش‌بینی عملکرد نیشکر در مزارع کشت و صنعت دهخدا (استان خوزستان) ارائه شد که مبتنی بر ادغام داده‌های ماهواره‌ای و اطلاعات زراعی است. داده‌های مورداستفاده شامل ۲۴۱۷ دبیزه از سال‌های زراعی ۱۳۹۶ تا ۱۴۰۳ بود که از سوابق مدیریتی مزرعه و تصاویر ماهواره‌ای سنتینل ۲ شامل شاخص‌های گیاهی NDVI و EVI استخراج شد. در گام نخست، مزارع با استفاده از الگوریتم K میانگین و بر اساس ویژگی‌های مدیریتی و عملکردی به چهار گروه متمایز خوشه‌بندی شدند. سپس برای بهبود دقت مدل، مجموعه‌ای از ویژگی‌های مهندسی‌شده مانند بهره‌وری آب و بهره‌وری کود تعریف گردید. در ادامه، دو مدل یادگیری ماشین شامل جنگل تصادفی و گرادیان بوستینگ برای پیش‌بینی عملکرد نهایی آموزش داده شد. برای ارزیابی مدل‌ها، ۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد برای آزمون مستقل در نظر گرفته شد و همچنین از اعتبارسنجی متقابل پنج‌لایه به‌منظور افزایش پایداری و عمومیت‌پذیری استفاده شد. نتایج نشان داد مدل گرادیان بوستینگ بهترین عملکرد را ارائه کرده و به ضریب تعیین ۰/۹۹۲۴ و ریشه میانگین مربعات خطا برابر ۱/۸۸ تن در هکتار دست‌یافته است. علاوه‌برآن، تحلیل اهمیت ویژگی‌ها نشان داد بهره‌وری آب با سهم ۸۷ درصد، اثرگذارترین عامل در پیش‌بینی عملکرد است. در مجموع، یافته‌ها تأیید می‌کنند که ترکیب داده‌های مدیریتی و ماهواره‌ای همراه با مهندسی ویژگی می‌تواند ابزاری دقیق و کارآمد برای پشتیبانی از تصمیم‌گیری‌های مکانی و مدیریت منابع در کشاورزی دقیق باشد.

نحوه استناد به این مقاله:

جادری، ف. و منجزی، ن. (۱۴۰۵). پیش‌بینی عملکرد نیشکر با رویکرد یکپارچه‌سازی شاخص‌های ماهواره‌ای Sentinel-2 و ویژگی‌های مدیریتی، با استفاده از الگوریتم‌های یادگیری ماشین گروهی. *نشریه پژوهش‌های مکانیک ماشین‌های کشاورزی*، ۴۰: ۳۱-۴۸.
https://dx.doi.org/10.22034/jrmam.2026.14989.762

مقدمه

نیشکر (*Saccharum officinarum*) به‌عنوان یکی از محصولات راهبُردی کشاورزی، نقشی محوری در اقتصاد جهانی، تولید شکر و تأمین اتانول و سوخت‌های زیستی ایفا می‌کند. بااین‌حال، پایداری زنجیره تأمین این محصول تنها به مدیریت مزرعه محدود نشده و تحت‌تأثیر عوامل متعددی از جمله مدیریت ریسک تأخیرات انتقال محصول به کارخانه قرار دارد که می‌تواند بر کارایی نهایی کل سامانه اثرگذار باشد (Afsharnia & Marzban, 2019). باتوجه‌به افزایش جمعیت و چالش‌های ناشی از تغییرات اقلیمی، دستیابی به پایداری در تولید نیشکر و بهینه‌سازی مدیریت منابع در استان خوزستان به‌عنوان قطب اصلی صنعت شکر ایران، به یک ضرورت اساسی تبدیل شده است (Seyed jalali *et al.,* 2021).

در این منطقه، حتی بهبودهای جزئی در دقت پیش‌بینی عملکرد پیش از برداشت، می‌تواند به صرفه‌جویی اقتصادی قابل‌توجه و مدیریت کارآمدتر نهاده‌ها منجر شود. در این میان، بهره‌گیری از مدل‌های یکپارچه محاسباتی و رویکردهای داده‌محور، راهکاری کلیدی برای تحلیل سامانه‌های پیچیده زراعی، کاهش چالش‌های محیطی و تضمین پایداری تولید در آینده است (Bhatt *et al.,* 2024; Nakhaeinejad *et al.,* 2022).

بااین‌حال، روش‌های سنتی تخمین عملکرد اغلب پرهزینه و زمان‌بر هستند. در سال‌های اخیر، فناوری‌های کشاورزی دقیق، به‌ویژه سنجش‌ازدور، راهکارهای نویدبخشی برای این چالش‌ها ارائه کرده‌اند. تصاویر ماهواره‌ای با وضوح بالا، مانند داده‌های سنتینل ۲ (Sentinel 2)، امکان پایش مستمر سلامت محصول را فراهم می‌کنند. اما تحلیل حجم عظیم این داده‌ها نیازمند ابزارهای محاسباتی پیشرفته است؛ لذا یادگیری ماشین به ابزاری استاندارد برای مدل‌سازی روابط پیچیده و غیرخطی بین داده‌های طیفی و عملکرد نهایی تبدیل شده است (Van Klompenburg *et al.,* 2020).

مروری بر پژوهش‌های پیشین

پیشینه تحقیق نشان می‌دهد که الگوریتم‌های یادگیری ماشین در ترکیب با داده‌های ماهواره‌ای نوری بسیار موفق عمل کرده‌اند. برای مثال، (Taravat *et al.,* 2024) با استفاده از مدل جنگل تصادفی و داده‌های چندزمانه سنتینل ۲ (Sentinel 2)، به توان پیش‌بینی قدرتمندی دست یافتند. در همین راستا، (Canata *et al.,* 2021) با استفاده از تصاویر با وضوح بالا و فن‌های یادگیری ماشین به‌دقت قابل‌توجهی در نقشه‌برداری عملکرد نیشکر دست یافتند. همچنین (Akbarian et al., 2024) کارایی این مدل‌ها را در سطح مزرعه تأیید کردند، درحالی‌که (Kaydani *et al.,* 2025) با تلفیق تصاویر ماهواره‌ای لندست (Landsat) و سنتینل (Sentinel) در اراضی نیشکر هفت‌تپه، پتانسیل بالای داده‌های سنجش از دوری را در تخمین دقیق عملکرد اثبات نمودند. در گامی فراتر، پژوهشگران به سراغ یکپارچه‌سازی منابع داده‌ای متنوع و پلتفرم‌های جایگزین رفته‌اند؛ (Das *et al.,* 2023) با ترکیب داده‌های نوری و راداری (SAR) و (Shendryk *et al.,* 2021) با ترکیب تصاویر ماهواره‌ای و داده‌های محیطی (دما و بارش)، دقت مدل‌ها را بهبود بخشیدند.

در بحث مدیریت سامانه‌های پیچیده زراعی،(Parchami-Araghi & Sadeghi-Lari, 2021) با تحلیل عدم قطعیت مدل‌سازی زراعی_هیدرولوژیکی در اراضی نیشکر، بر ضرورت پایش دقیق اثرات متقابل مدیریت منابع آب و خاک تأکید کردند. یافته‌های (de Oliveira Maia *et al.,* 2023) نیز نشان داد که شاخص‌های گیاهی ابزاری قدرتمند برای پهنه‌بندی مدیریتی در مزارع نیشکر هستند، رویکردی که در سایر علوم محیطی نیز سابقه موفقی داشته است، به‌طوری که (Karami & Esmaili-Rineh, 2020) نشان دادند تلفیق روش‌های خوشه‌بندی K-means با مدل‌های الگوریتمی، کارایی بالایی در تفکیک الگوهای پیچیده مکانی دارد. جدای از پلتفرم‌های ماهواره‌ای، پهپادها نیز برای جمع‌آوری داده‌هایی با وضوح مکانی فوق‌العاده بالا مورداستفاده قرار گرفته‌اند، به‌عنوان نمونه (Poudyal *et al.,* 2022) از تصاویر فراطیفی مبتنی بر پهپاد برای پیش‌بینی دقیق عملکرد استفاده کردند.

در همین رابطه، (Mariadass *et al.,* 2022) با به‌کارگیری الگوریتم XGBoost و روش‌های تفسیری، دقت پیش‌بینی عملکرد را در شرایط اراضی ناهمگن ارتقا دادند. قابلیت تعمیم این روش‌های هوشمند به محصولات دیگر نیز مانند مطالعه (Shahhosseini *et al.,* 2021) که یادگیری ماشین و مدل‌سازی زراعی را برای پیش‌بینی عملکرد ذرت به کار بردند، تأیید شده است. با پیشرفت ابزارهای پایش، (Singla *et al.,* 2021) توانایی مدل‌های هوشمند یادگیری جمعی را در تفکیک پاسخ‌های طیفی مزارع نیشکر بر اساس چرخه‌های زراعی (کشت اول در برابر بازرویی) موردتأیید قرار دادند.

اخیراً، رویکرد پژوهشگران از تمرکز صرف بر یافتن منابع داده جدید، به سمت مدیریت هوشمندانه‌تر ویژگی‌های موجود معطوف شده است. همان‌طور که در مرور سامان‌مند (de França e Silva *et al.*, 2024) موردتأکید قرار گرفته، عملکرد مدل‌های تجربی به‌شدت به کیفیت و قدرت توضیحی متغیرهای ورودی وابسته است. در همین رابطه، (Soltani-Kazemi *et al.*, 2024) با مقایسه الگوریتم‌های مختلف یادگیری ماشین برای برآورد رطوبت غلاف نیشکر، نشان دادند که مدل‌های غیرخطی توانمندی بالاتری در رهگیری عامل‌های بیوفیزیکی گیاه در طول دوره رشد دارند. روش‌هایی مانند انتخاب راهبُردی ویژگی‌ها برای بهینه‌سازی عملکرد مدل ضروری هستند، برای نمونه (Charoen-Ung & Mittrapiyanuruk, 2018) بر انتخاب ویژگی پیش‌رونده تمرکز کردند تا مؤثرترین متغیرها را برای مدل جنگل تصادفی شناسایی کنند.

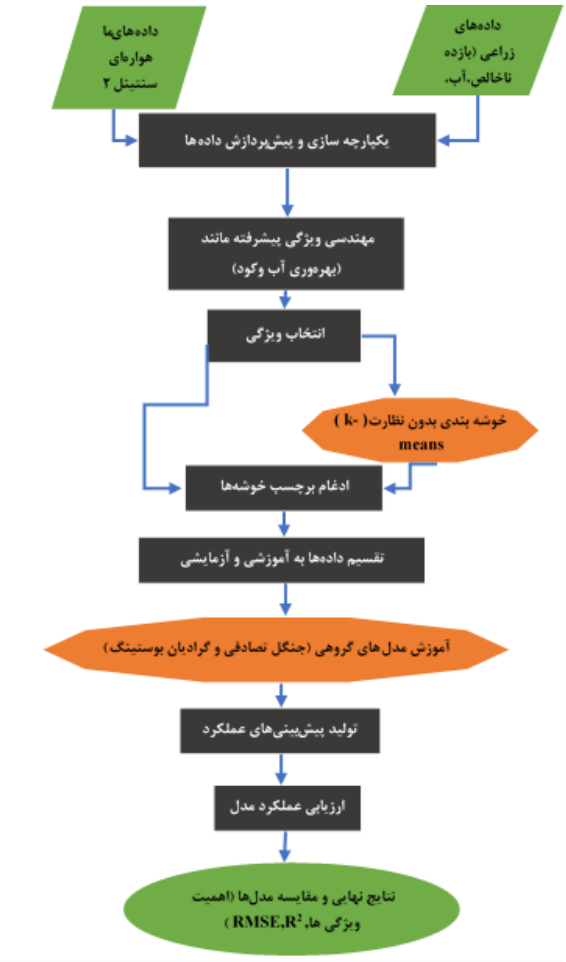
همچنین (Mirzaei *et al.*, 2021) با استفاده از روش بردار پشتیبان بر روی داده‌های طیفی، موفق به برآورد غیرمخرب شاخص‌های حیاتی برگ شدند. بااین‌حال، اکثر مطالعات موجود همچنان عمدتاً بر شاخص‌های گیاهی استاندارد و از پیش تعریف‌شده متکی هستند، در همین راستا، (Parashar *et al.*, 2024) اثبات کردند که این شاخص‌های طیفی سنتی به‌تنهایی قادر به انعکاس دقیق کارایی مدیریت زراعی و تنش‌های موضعی مزارع نبوده و در فرایند واسنجی با محدودیت مواجه می‌شوند. ازاین‌رو، پتانسیل ایجاد ویژگی‌های مهندسی‌شده نوآورانه از طریق یکپارچه‌سازی هم‌زمان داده‌های مدیریتی و ماهواره‌ای، کماکان یک خلأ تحقیقاتی حیاتی محسوب می‌شود که مبنای توسعه چارچوب پیشنهادی در این پژوهش قرار گرفته است.

نوآوری و هدف پژوهش

با وجود پیشرفت‌های مذکور، همچنان شکاف تحقیقاتی در زمینه یکپارچه‌سازی داده‌های دقیق مدیریتی (مانند بهره‌وری منابع) با داده‌های ماهواره‌ای وجود دارد. پژوهش حاضر باهدف پر کردن این خلاء، چارچوبی جامع برای پیش‌بینی عملکرد نیشکر ارائه می‌دهد. نوآوری اصلی این تحقیق در دو جنبه نهفته است: نخست، مدیریت ناهمگنی فضایی مزارع با استفاده از الگوریتم خوشه‌بندی K-means برای دسته‌بندی مزارع بر اساس الگوهای مدیریتی؛ و دوم، معرفی ویژگی‌های مهندسی‌شده جدید (مانند بهره‌وری آب و کود) که تاثیر مستقیم مدیریت مزرعه را در مدل لحاظ می‌کنند. این رویکرد فراتر از تکیه صرف بر شاخص‌های گیاهی سنتی عمل کرده و با مقایسه مدل‌های گروهی جنگل تصادفی و گرادیان بوستینگ، به دنبال دستیابی به حداکثر دقت پیش‌بینی در شرایط عملیاتی است.

مواد و روش‌ها

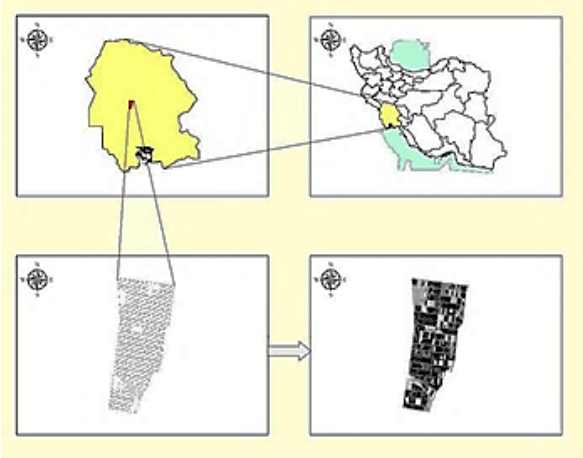
چارچوب پیشنهادی برای پیش‌بینی عملکرد نیشکر، یک فرایند چندمرحله‌ای و داده‌محور است که در (شکل ۱) نمایش داده شده است. این فرایند با جمع‌آوری و یکپارچه‌سازی داده‌های چند منبعی آغاز می‌شود. در گام بعدی، با مهندسی ویژگی‌ها، تلاش می‌کنیم تا اثرات مدیریت زراعی را به‌صورت صریح در مدل لحاظ کرده و اطلاعات پنهان را استخراج کنیم. سپس، از خوشه‌بندی بدون نظارت برای طبقه‌بندی مزارع و مدیریت ناهمگنی فضایی استفاده شده است. در نهایت، پیش‌بینی دقیق عملکرد از طریق مدل‌های گروهی نظارت‌شده انجام می‌شود. جزئیات هر مرحله در ادامه تشریح شده است.



شکل۱- معماری کلی چارچوب پیشنهادی برای پیش‌بینی عملکرد

منطقه مورد مطالعه

این پژوهش در مزارع نیشکر متعلق به شرکت کشت و صنعت
این پژوهش در مزارع نیشکر متعلق به شرکت کشت و صنعت
نیشکر دهخدا واقع در استان خوزستان در بازه سال‌های زراعی ۱۳۹۶ تا ۱۴۰۳ انجام شده است (شکل ۲). این کشت و صنعت در شمال شهرستان اهواز، در محدوده‌ای بین جاده اندیمشک و رودخانه کارون و در کیلومتر ۲۳ جاده دغاغله قرار دارد. مساحت خالص اراضی زیرکشت این شرکت حدود ۱۱۳۶۹ هکتار بوده و ظرفیت اسمی تولید آن به‌ترتیب شامل ۱۰۰۰۰۰ تن شکر خام، ۱۷۵۰۰۰ تن شکر سفید، ۳۰۰۰۰ تا ۳۸۰۰۰ تن ملاس و فرآورده‌های جانبی آن و حدود ۳۳۰۰۰۰ تن باگاس در سال است.



شکل ۲- موقعیت کشت و صنعت دهخدا در استان خوزستان

توصیف مجموعه داده

مجموعه داده مورداستفاده در این پژوهش، از طریق یکپارچه‌سازی دو منبع داده مجزا تدوین گردید. منبع نخست، شامل داده‌های زراعی و مدیریتی است که ۲۴۱۷ دبیزه از سوابق قطعات مزارع در بازه زمانی سال‌های زراعی ۱۳۹۶ تا ۱۴۰۳ را در بر می‌گیرد. متغیرهای کلیدی استخراج‌شده از این سوابق عبارتند از: عملکرد ناخالص نهایی (تن در هکتار)، میزان آب مصرفی (مترمکعب در هکتار)، کل کود مصرفی (کیلوگرم در هکتار) و شوری خاک (EC) به‌منظور اطمینان از تعمیم‌پذیری و پوشش کامل شرایط عملیاتی، نمونه‌ها شامل تمامی چرخه‌های زراعی (از کشت اول تا آخرین راتون‌های مدیریتی) بوده و ارقام غالب مورد بررسی شامل CP69-1062، CP57-614، SP70-114،CP48-103 و IRC99-02 می‌باشند. منبع دوم، داده‌های ماهواره‌ای است. در این بخش، از تصاویر چندطیفی سنتینل ۲ با سطح پردازش Level-2A (تصحیح اتمسفری شده) استفاده شد. این ماهواره با وضوح مکانی بالا (۱۰متر در باندهای مرئی و مادون‌قرمز نزدیک) و دوره بازگشت کوتاه (تقریباً ۵ روز)، ابزاری ایده‌آل برای پایش کشاورزی در مقیاس وسیع محسوب می‌شود. برای ثبت حداکثر زیست‌توده گیاهی و اطلاعات مرتبط، تصاویر مربوط به اوج فصل رشد (ماه‌های شهریور تا آذر) برای هر سال زراعی انتخاب شدند.

شاخص‌های گیاهی و پیش‌پردازش

جهت محاسبه شاخص‌های گیاهی مورداستفاده در پژوهش (شاخص پوشش گیاهی نرمال شده تفاضلی NDVI و شاخص پوشش گیاهی بهبود یافته EVI)، از روابط استاندارد (۱) و (۲) استفاده گردید.

$$NDVI = \frac{R_{NIR} - R_{Red}}{R_{NIR} + R_{Red}} \quad (۱)$$

$$EVI = G \times \frac{R_{NIR} - R_{Red}}{R_{NIR} + C1 \times R_{Red} - C2 \times R_{Blue} + L} \quad (۲)$$

که در این روابط، R بیانگر بازتابش طیفی در باندهای قرمز (Red)، آبی (Blue) و مادون‌قرمز نزدیک (NIR) است. ضرایب استاندارد رابطه EVI شامل L: ۱، C_1: ۶ ، C_2: ۷/۵ و G : ۲/۵ در نظر گرفته شدند.

فرایند انتخاب ویژگی منجر به گزینش نهایی این دو شاخص گردید؛ به‌طوری که شاخص‌های دارای هم‌بستگی بالا جهت جلوگیری از پدیده هم‌خطی چندگانه حذف شدند. همچنین، پیش‌پردازش داده‌ها شامل حذف داده‌های پرت و خطاهای اندازه‌گیری (به‌ویژه در شاخص EVI) انجام یافت. در نهایت، مقادیر میانگین شاخص‌ها برای دوره رشد محاسبه و بر اساس کد شناسایی مزرعه و سال زراعی با پایگاه‌داده مدیریتی ادغام شد.

پیش‌پردازش و مهندسی ویژگی

پیش از آموزش مدل، مرحله پیش‌پردازش بر روی مجموعه‌داده انجام گرفت. نخست، باتوجه‌به ناچیز بودن مقادیر گمشده (کمتر از ۳ درصد کل پایگاه‌داده)، این مقادیر با استفاده از میانگین ستون مربوطه جایگزین شدند تا از سوگیری ویژگی‌های مدیریتی خاص مزارع جلوگیری شود. سپس، برای جلوگیری از سوگیری مدل به سمت متغیرهای با مقیاس بزرگ‌تر، تمام ویژگی‌ها با استفاده از تابع StandardScaler در کتابخانه Scikit-learn استانداردسازی شدند تا دارای میانگین

صفر و واریانس واحد باشند. مهندسی ویژگی که نوآوری اصلی این تحقیق را تشکیل می‌دهد، در گام بعدی انجام شد. به‌جای تکیه صرف بر شاخص‌های گیاهی استاندارد، ویژگی‌های جدیدی باهدف مدل‌سازی مستقیم کارایی مدیریت زراعی و استخراج اطلاعات سطح بالاتر خلق شدند. این رویکرد، مشابه مفهوم مهندسی ویژگی که توسط (Barbosa *et al.*, 2023) معرفی شده، به دنبال استخراج اطلاعات سطح بالاتر از داده‌های خام است. پنج ویژگی مهندسی‌شده اصلی به شرح زیر تعریف و محاسبه شدند:

بهره‌وری آب: نشان می‌دهد که به‌ازای هر واحد آب مصرفی، چه میزان محصول تولید شده است و نماینده کارایی آبیاری است.

بهره‌وری کود: این ویژگی کارایی مصرف کود را مدل‌سازی می‌کند.

شاخص سلامت پوشش گیاهی: این شاخص یک دید کلی و ترکیبی از سلامت گیاه ارائه می‌دهد.

نسبت آبیاری به کود: این نسبت، راهبُرد کلی مدیریت مزرعه در تخصیص منابع را نشان می‌دهد.

اثر متقابل شوری خاک و پوشش گیاهی: این ویژگی به دنبال مدل‌سازی این فرضیه است که تأثیر شوری خاک بر سلامت گیاه ممکن است غیرخطی باشد.

پس از خلق ویژگی‌ها، از الگوریتم SelectKBest با آزمون آماری F-regression برای انتخاب ۲۵ ویژگی برتر استفاده شد. این رویکرد برای انتخاب متغیرهای حساس، مشابه روش‌های پیشنهادی توسط (Zhang *et al.*, 2025) به کاهش ابعاد و افزایش پایداری مدل کمک می‌کند.

چارچوب پیش‌بینی خوشه‌بندی و رگرسیون

مرحله ۱: خوشه‌بندی بدون نظارت مزارع با K-means برای شناسایی الگوهای فضایی و گروه‌بندی مزارع با مشخصات مشابه، از الگوریتم خوشه‌بندی K-means استفاده شد. این الگوریتم با کمینه‌سازی مجموع مربعات فاصله درون‌خوشه‌ای (اینرسی) عمل می‌کند که تابع هدف آن به‌صورت رابطه (۳) تعریف می‌شود.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \quad (۳)$$

که در آن: J تابع هدف، k تعداد خوشه‌ها (که با روش آرنج برابر ۴ تعیین شد)، و cj مرکز خوشه j است. برچسب خوشه حاصل به‌عنوان یک ویژگی جدید به مجموعه داده اضافه شد تا مدل‌های بعدی بتوانند از این اطلاعات ساختاری بهره‌مند شوند.

مرحله ۲: پیش‌بینی نظارت‌شده عملکرد

در این مرحله از پژوهش، دو مدل گروهی قدرتمند یعنی جنگل تصادفی و گرادیان بوستینگ، برای پیش‌بینی عملکرد نهایی مورداستفاده و توسعه قرار گرفتند. مدل گرادیان بوستینگ که بهترین عملکرد را از خود نشان داد، یک مدل افزایشی است که بر اساس منطق تقویت کار می‌کند. این مدل، درخت‌های تصمیم ضعیف را به‌صورت متوالی به ساختار خود اضافه می‌کند و هر درخت جدید، تلاش می‌کند تا خطاهای باقیمانده مدل قبلی را اصلاح و کاهش دهد. این الگوریتم، برای بهینه‌سازی مستقیم تابع هزینه، از روش گرادیان کاهشی در فضای توابع استفاده می‌کند. مدل نهایی، پس از m مرحله تکرار، با استفاده از رابطه ۴ نهایی می‌شود.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (۴)$$

که در آن: Fm(x) مدل نهایی، hm(x) درخت تصمیم جدید، و γm نرخ یادگیری است که اندازه گام در هر مرحله را کنترل می‌کند و به جلوگیری از بیش برازش کمک می‌کند.

شایان ذکر است که در این پژوهش، به‌منظور بهره‌گیری حداکثری از پتانسیل پهنه‌بندی اراضی، الگوریتم خوشه‌بندی K-means بر روی کل پایگاه‌داده اعمال گردید تا مرزهای مدیریتی مزارع مشخص شوند. بااین‌حال، از منظر جداسازی مطلق خط لوله یادگیری ماشین و جلوگیری از هرگونه سوگیری احتمالی در فاز پیش‌پردازش، فرایند استانداردسازی و تحلیل‌ها به گونه‌ای مدیریت شد که واریانس و میانگین کلی داده‌ها به‌عنوان یک فرض ثابت منطقه‌ای لحاظ شود، اگرچه پیشنهاد متدولوژیک برای ساختارهای کاملاً مستقل، تفکیک اولیه داده‌ها پیش از هرگونه پردازش طیفی است.

ابزارها، آموزش مدل و معیارهای ارزیابی

تمام مراحل تحلیل داده و مدل‌سازی با استفاده از زبان برنامه‌نویسی پایتون (نسخه ۳.۱۳.۲) و کتابخانه‌های تخصصی آن صورت پذیرفت. این فرایند شامل به‌کارگیری Pandas برای مدیریت و سازماندهی داده‌ها، Scikit-learn برای پیاده‌سازی مدل‌های یادگیری ماشین و Matplotlib برای بصری‌سازی نتایج بود.

مجموعه داده به نسبت ۸۰ درصد آموزشی و ۲۰ درصد آزمایشی با روش طبقه‌بندی‌شده بر اساس برچسب خوشه‌ها تقسیم شد. برای ارزیابی پایداری مدل، از اعتبارسنجی متقابل ۵لایه استفاده شد. در نهایت، عملکرد مدل با استفاده از سه معیار استاندارد آماری که در روابط ۵ تا ۷ تعریف شده‌اند مورد سنجش و ارزیابی قرارگرفت.

۱) ضریب تعیین (R^2)

$$R^2 = \frac{(\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y}))^2}{\sum_{i=1}^{n}(xi-\bar{x})^2 \sum_{i=1}^{n}(yi-\bar{y})^2} \quad (۵)$$

۲) ریشه میانگین مربعات خطا (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(xi-yi)^2}{N}} \quad (۶)$$

۳) میانگین قدر مطلق خطا (MAE)

$$MAE = \frac{\sum_{i=1}^{N}|xi-yi|}{N} \quad (۷)$$

که در آن: xi مقادیر پیش‌بینی شده، $\bar{x}$ میانگین مقادیر پیش‌بینی شده، yi مقادیر مشاهده شده (واقعی)، $\bar{y}$ میانگین مقادیر مشاهده شده و N تعدادت مشاهدات است.

نتایج و بحث

این بخش به ارائه، تحلیل و بحث درباره یافته‌های تجربی پژوهش می‌پردازد. ساختار این بخش به این صورت است که ابتدا، آمار توصیفی داده‌ها و نتایج خوشه‌بندی مزارع گزارش می‌شود. سپس، عملکرد مدل‌های پیش‌بینی به تفصیل ارزیابی و مقایسه خواهد شد و در نهایت، تحلیل اهمیت ویژگی‌ها و بررسی عملکرد مدل در خوشه‌های مختلف مورد بحث و بررسی قرار می‌گیرد.

تحلیل آماری داده‌ها و نتایج خوشه‌بندی

مجموعه داده نهایی، ۲۴۱۷ دبیزه از مزارع نیشکر را در بر می‌گیرد. جدول ۱ آمار توصیفی متغیرهای کلیدی را نمایش می‌دهد.

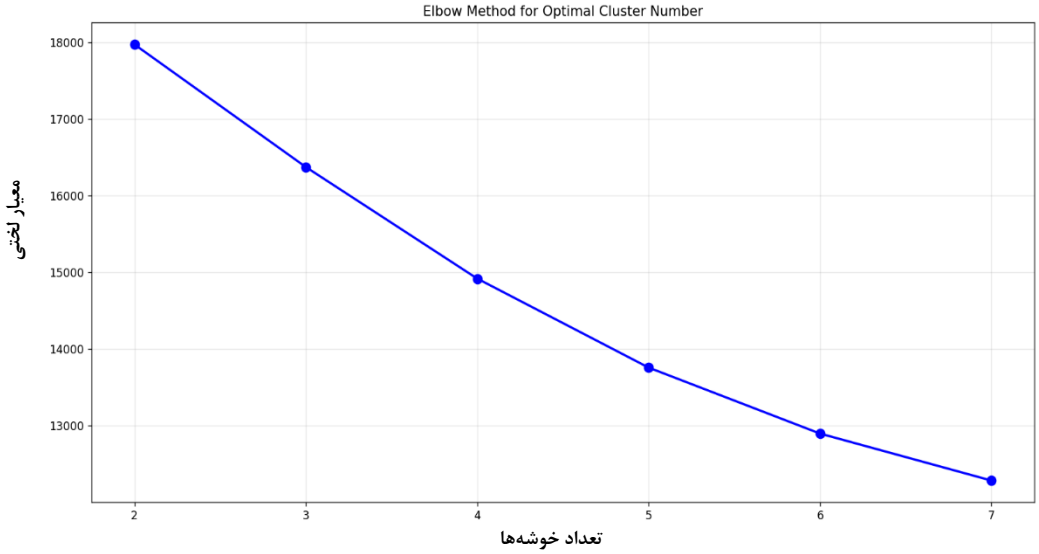
جدول ۱-آمار توصیفی متغیرهای کلیدی

متغیر	میانگین	انحراف معیار	حداقل	حداکثر
بازده ناخالص نهایی (تن در هکتار)	۶۷/۱۶	۲۲/۱۷	۱۴/۵۸	۱۴۷/۱۳
مصرف آب (مترمکعب در هکتار)	۱۴۲۰/۹۵	۱۲۹/۰۷	۱۰۰۴/۷۸	۱۹۰۰/۰۰
کل کود مصرفی(کیلوگرم در هکتار)	۳۴۸/۸۶	۴۷/۷۰	۱۶۰/۲۷	۵۵۳/۳۳
شوری خاک	۵/۴۷	۲/۱۱	۱/۱۰	۱۶/۴۰

بررسی آمار توصیفی نشان می‌دهد که متغیر هدف (بازده نهایی) با میانگین ۶۷/۱۶ تن در هکتار، دارای انحراف معیار قابل‌توجه ۲۲/۱۷ تن در هکتار است. این دامنه تغییرات گسترده، حاکی از تأثیر شدید عوامل محیطی و مدیریتی بر عملکرد نیشکر در منطقه مورد مطالعه است. وجود این پراکندگی، از یک سو پیچیدگی ذاتی مسئله پیش‌بینی را نشان می‌دهد و از سوی دیگر، ضرورت استفاده از مدل‌های غیرخطی یادگیری ماشین و فن‌های پیش‌پردازش مانند خوشه‌بندی را تایید می‌کند؛ چرا که مدل‌های رگرسیون خطی ساده معمولاً در مواجهه با چنین واریانس بالایی دچار خطای سامان‌مند می‌شوند. علاوه بر این، ضریب تغییرات بالا در متغیرهای میزان آب مصرفی و شوری خاک نشان‌دهنده ناهمگنی شرایط زراعی در قطعات مختلف است. این موضوع توجیه‌کننده استفاده از الگوریتم K-means در این پژوهش است، زیرا با دسته‌بندی مزارع به گروه‌های همگن‌تر، می‌توان اثر این نوسانات شدید مدیریتی را در مدل پیش‌بینی تعدیل کرد. در واقع، خوشه‌بندی اجازه می‌دهد تا مدل برای هر گروه از مزارع با شرایط مشابه، وزن‌دهی دقیق‌تری به ویژگی‌ها اختصاص دهد.

تحلیل خوشه‌بندی و ناهمگنی زراعی(K-means)

به‌منظور کنترل ناهمگنی ساختاری در بانک اطلاعاتی مزارع، الگوریتم خوشه‌بندی K-means به‌کار گرفته شد. مطابق با روش آرنج (شکل ۳)، نقطه عطف نمودار در K=۴ تعیین گردید که بیانگر بهینه‌ترین تعداد گروه‌بندی برای پوشش حداکثری واریانس داده‌هاست. ویژگی‌های تفکیکی این چهار خوشه در جدول ۲ ارائه شده است.



شکل ۳- تعیین تعداد بهینه خوشه ها (K=۴) با استفاده از روش آرنج

تحلیل داده‌های جدول ۲ نشان می‌دهد که الگوریتم K-means به‌خوبی توانسته است الگوهای متناقض مدیریتی و محیطی حاکم بر مزارع را تفکیک کند.

خوشه ۲ (پتانسیل بالا): با عملکرد ۷۶/۷۸ تن، بهترین وضعیت را دارد. این مزارع در خاک‌های با کمترین شوری (۴/۵۹) قرار داشته و بیشترین میزان کود (۳۳۶/۲۶کیلوگرم) را دریافت کرده‌اند که نشان‌دهنده پاسخ‌دهی حداکثری گیاه به نهاده در غیاب تنش است.

خوشه ۴ (پایداری در تنش): با عملکرد ۷۱/۵۴ تن در رتبه دوم قرار دارد. نکته تحلیلی این خوشه، حفظ عملکرد مطلوب علیرغم شوری بالا (۵/۹۲) است که می‌تواند ناشی از انتخاب ارقام مقاوم‌تر یا مدیریت دقیق‌تر دوره‌های آبیاری باشد.

خوشه ۱ (تنش شوری و هدررفت منابع): علیرغم مصرف بالای آب (۱۴۳۳/۶۲ مترمکعب)، عملکرد در سطح ۶۳/۱۳ تن متوقف شده است. این امر به دلیل شوری بحرانی خاک (۶/۱۰) رخ داده که با کاهش پتانسیل اسمزی، کارایی مصرف آب را به‌شدت افت داده است.

خوشه ۳ (ضعیف): با کمترین عملکرد (۵۸/۴۷تن)، بازتاب‌دهنده مزارعی است که هم‌زمان با تنش محیطی و کمترین میزان تخصیص نهاده‌ها (کود و آب) مواجه بوده‌اند.

این طبقه‌بندی با کنترل ناهمگنی فضایی که پیش‌تر در انحراف معیار بالای عملکرد (۲۲/۱۷) در جدول ۱ مشاهده شده بود، اختلال داده‌ها را کاهش داده و دقت مدل‌های یادگیری ماشین را در مراحل بعدی تضمین کرده است.

جدول ۲- ویژگی‌های خوشه‌های مزرعه شناسایی شده

خوشه	تعداد نمونه‌ها	میانگین عملکرد (تن در هکتار)	میانگین آب (مترمکعب در هکتار)	میانگین کود (کیلوگرم در هکتار)	میانگین شوری خاک
۱	۷۶۱	۶۳/۱۳	۱۴۳۳/۶۲	۳۵۱/۸۳	۶/۱۰
۲	۷۲۱	۷۶/۷۸	۱۳۹۹/۷۵	۳۶۶/۲۶	۴/۵۹
۳	۶۰۹	۵۸/۴۷	۱۴۳۳/۱۲	۳۳۲/۵۴	۵/۵۰
۴	۳۲۶	۷۱/۵۴	۱۴۱۵/۵۰	۳۳۳/۹۶	۵/۹۲

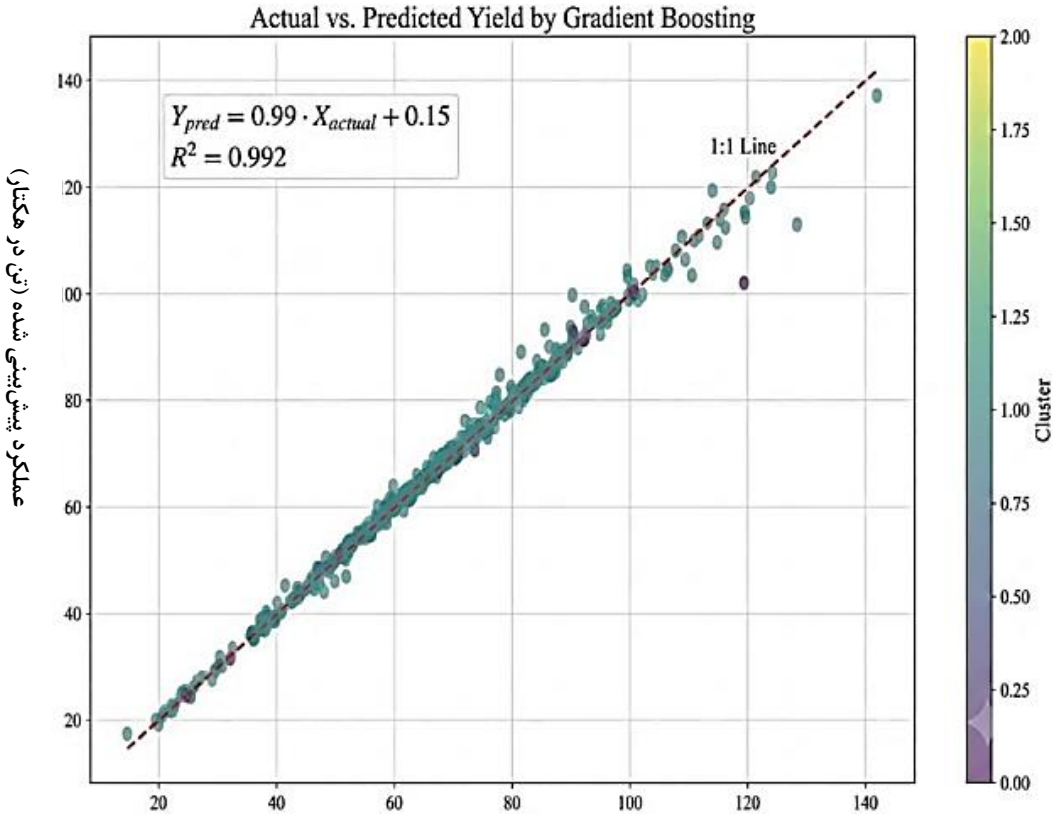
ارزیابی و مقایسه عملکرد مدل‌های رگرسیونی

به‌منظور پیش‌بینی دقیق عملکرد نیشکر، کارایی دو الگوریتم یادگیری ماشین گروهی شامل جنگل تصادفی و افزایش گرادیان بر روی داده‌های آزمون (تست) مورد ارزیابی قرار گرفت. معیارهای آماری مقایسه مدل‌ها شامل ضریب تبیین (R^2)، ریشه میانگین مربعات خطا (RMSE) و میانگین خطای مطلق (MAE) در جدول ۳ ارائه شده است.

جدول ۳- مقایسه عملکرد مدل‌های یادگیری ماشین

مدل	R^2	RMSE	MAE
جنگل تصادفی	۰/۹۸۰۶	۳/۰۲۲۶	۱/۹۳۷۸
گرادیان بوستینگ	۰/۹۹۲۴	۱/۸۸۹۵	۱/۰۷۹۸

مطابق با یافته‌های جدول ۳، هر دو مدل مبتنی بر درخت از دقت و توان پیش‌بینی بسیار بالایی در تخمین عملکرد برخوردارند. بااین‌حال، مدل گرادیان بوستینگ با تخصیص بالاترین ضریب تبیین (۰/۹۹۲۴) و کمترین میزان خطا ۱/۸۸۹۵ تن در هکتار(، عملکرد برتری نسبت به جنگل تصادفی نشان می‌دهد. نمودار پراکنش مقادیر واقعی در برابر پیش‌بینی‌شده (شکل ۴) نیز انطباق بسیار بالای داده‌ها حول خط ۱:۱ (۰/۹۹۲۴) را تأیید می‌کند.



شکل ۴- همبستگی پیش‌بینی‌شده توسط گرادیان بوستینگ

دلیل اصلی دستیابی به این دقت متمایز و پیشرفته، عملکرد متوالی الگوریتم گرادیان بوستینگ است، این مدل در هر گام، درخت‌های جدید را برای اصلاح و جبران خطای (باقی‌مانده‌های) درختان قبلی آموزش می‌دهد، درحالی‌که جنگل تصادفی درخت‌ها را به‌صورت موازی و مستقل می‌سازد. از منظر مهندسی سامانه‌ها، این همگرایی فوق‌العاده بالا مستقیماً به راهبُرد دومرحله‌ای این پژوهش یعنی «خوشه‌بندی پیش از رگرسیون» مرتبط است. از آنجا که الگوریتم K-means در مرحله قبل (جدول ۲) موفق شد نویز ناشی از ناهمگنی‌های شدید زراعی (مانند نوسانات تداخل شوری و مدیریت نهاده‌ها) را کنترل و مزارع را به ۴ خوشه با ویژگی‌های درون‌خوشه‌ایِ همگن تفکیک کند، متغیرهای مهندسی‌شده ورودی (به‌ویژه ویژگی‌های کلیدی بهره‌وری آب و کود) با بالاترین ظرفیت توضیحی در اختیار مدل رگرسیونی قرار گرفتند. در نتیجه، الگوریتم گرادیان بوستینگ توانسته است روابط غیرخطی پیچیده بین شاخص‌های طیفی ماهواره‌ای و فاکتورهای مدیریتی را با کمترین لغزش آماری کشف کند.

تحلیل اهمیت ویژگی‌ها و اولویت‌بندی عوامل مدیریتی و ماهواره‌ای

یافته‌های حاصل از ارزیابی اهمیت ویژگی‌ها در مدل برتر گرادیان بوستینگ نشان داد که ساختار غیرخطی این الگوریتم به‌خوبی توانسته است روابط پیچیده میان متغیرهای فرایندی و عملکرد نیشکر را استخراج کند (جدول ۴). بر اساس نتایج، شاخص‌های مهندسی‌شده مرتبط با مدیریت آب و نهاده‌های

شیمیایی، بالاترین سهم را در کاهش انحراف معیار خروجی و تبیین تغییرات عملکرد ناخالص اراضی به خود اختصاص داده‌اند. کارایی مصرف آب (با وزن اهمیت ۰/۰۸۷) و کارایی مصرف کود (۰/۰۸۶) به‌عنوان کلیدی‌ترین محرک‌های پیش‌بینی مدل شناسایی شدند. این امر نشان‌دهنده آن است که مدل هوشمند فراتر از تکیه بر متغیرهای منفرد، نسبت‌های تعاملی میان مصارف اراضی و خروجی بیوماس را به‌عنوان امضای اصلی رفتار مزارع شناسایی می‌کند.

رتبه سوم اهمیت به متغیر کل کود مصرفی (۰/۰۳۲) اختصاص‌یافته است که گویای وزن بالای مدیریت تغذیه در پایداری عملکرد است. بر خلاف الگوهای خطی سنتی، مدل‌های مبتنی بر درخت با ترکیب متغیرهای اتمسفری، زمانی و شاخص‌های گیاهی نوری استخراج‌شده از سنجش‌ازدور (مانند سری زمانی NDVI و EVI)، قادرند اثرات اشباع سنجه‌های ماهواره‌ای را از طریق تلفیق با فاکتورهای فیزیکی خاک (مانند EC) خنثی کنند.

جدول ۴- لیست ویژگی های برتر بر اساس اهمیت

٪ اهمیت	ویژگی
۰/۸۷۰	**کارایی مصرف آب**
۰/۰۸۶	**کارایی کود**
۰/۰۳۲	**کل کود مصرفی**

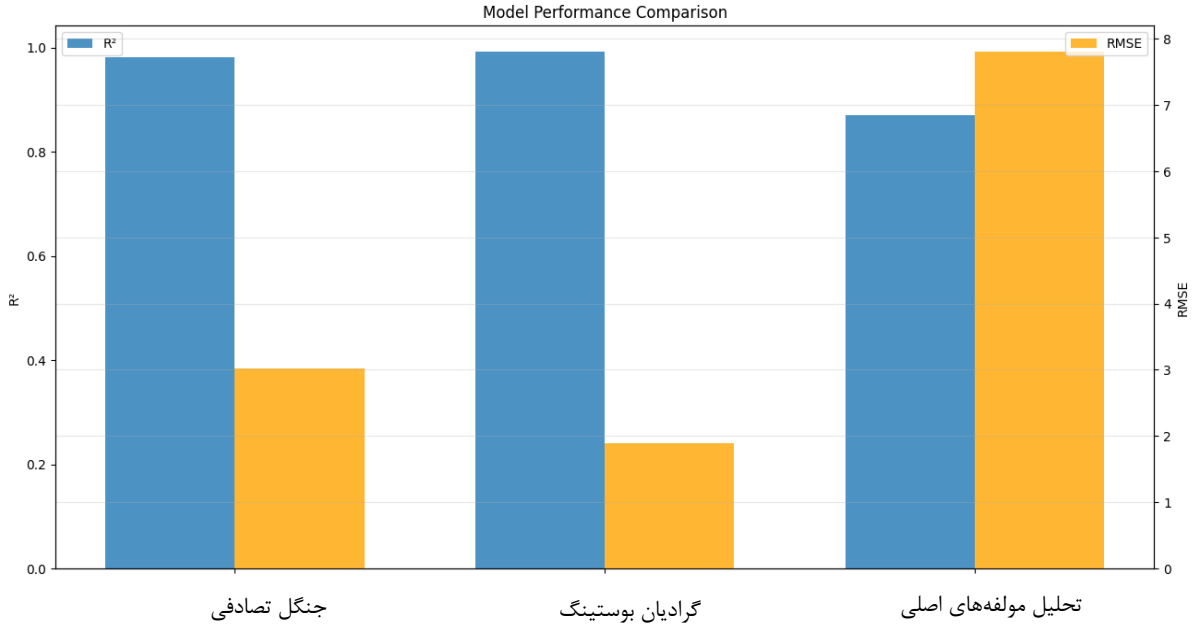
ساختار مقایسه‌ای عملکرد الگوهای توسعه‌یافته در این تحقیق (شکل ۵) نشان می‌دهد که یکپارچه‌سازی هوشمند پایگاه داده‌های مدیریتی زمینی با ویژگی‌های چندزمانی سنجش‌ازدور، به شکل پایداری خطا را کاهش و توانایی تعمیم مدل را بهبود بخشیده است. این یافته با مطالعات قبلی از جمله (Li *et al.*, 2024)که صرفاً بر روی قابلیت‌های داده‌های راداری (SAR) بدون در نظر گرفتن شاخص‌های مدیریت زراعی متمرکز بوده‌اند، مغایرت دارد. نتایج این بخش اثبات می‌کند که در سامانه‌های کشت متمرکز زراعی نظیر کشت نیشکر، فاکتورهای دینامیک مدیریتی (آب و کود) وزن بالاتری نسبت به فاکتورهای صرفاً مشاهداتی ماهواره‌ای در کالیبراسیون مدل‌های یادگیری ماشین دارند.

ارزیابی پتانسیل زراعی و تنوع عملکردی ارقام نیشکر

بررسی سنجه‌های آمار توصیفی مرتبط با ارقام مختلف نیشکر کشت‌شده در منطقه مورد مطالعه (جدول ۵ و شکل ۶)، گویای وجود یک ناهمگنی ژنتیکی ملموس و واریانس ساختاری بالا در ظرفیت بیوماس اراضی است. بر اساس نتایج، دامنه تغییرات میانگین بازده ناخالص نهایی در میان ارقام پنج‌گانه، از کمترین میزان بهره‌وری زراعی در رقم CP57-614 (با میانگین ۵۰/۰۸ تن بر هکتار) تا بالاترین سقف عملکرد مزارع در رقم-IRC99 02(با میانگین ۸۴/۵۵ تن بر هکتار) نوسان دارد. این تفاوت دامنه‌دار تایید می‌کند که ترجیحات ژنتیکی پتانسیل رقم، یکی از کلیدی‌ترین مولفه‌های مدیریتی در کالیبراسیون و دست‌یابی به بازدهی پایدار محسوب می‌شود.

واکاوی عمیق‌تر شاخص‌های پراکندگی ارقام، تضاد رفتاری و زراعی جالب‌توجهی را آشکار می‌سازد. رقم پربازده IRC99-02 با وجود داشتن بالاترین میانگین عملکرد، انحراف معیار بسیار بالایی (۳۱/۱۵) را ثبت کرده است که گویای حساسیت شدید فیزیولوژیک، ریسک‌پذیری زراعی و واکنش نوسانی این ژنوتیپ به تغییرات فاکتورهای محیطی و مدیریتی اراضی است. در نقطه مقابل، رقم کم‌بازده CP57-614 با ثبت کمترین انحراف معیار در میان تمامی ارقام (۱۴/۷۶)، پایداری رفتاری فوق‌العاده یکنواخت و یکدستی را در سطح مزارع نشان می‌دهد. همچنین رقم پرفراوانی نظیر CP69-1062 با حجم نمونه وسیع ۱۱۵۰ تایی، انحراف معیار کنترل‌شده‌تری (۱۹/۹۴) نسبت به رقم پربازده از خود بر جای گذاشته که نشان‌دهنده سازگاری عمومی مناسب این ژنوتیپ در مواجهه با نوسانات موضعی شوری خاک یا توزیع آب است.

از منظر محاسبات هوشمند، وجود این ناهمگنی شدید در واریانس‌ها به همراه عدم توازن شدید در فراوانی داده‌ها (مانند ۱۱۵۰ نمونه در برابر ۱۱۲ نمونه)، مدل‌های رگرسیون کلاسیک را دچار خطای سوگیری جدی می‌کند. بااین‌حال، متدولوژی الگوریتم کلاسترینگ توافقی به‌کاررفته در این تحقیق، از طریق گروه‌بندی مزارع در خوشه‌های مدیریت همگن، اثر این عدم توازن چگالی را خنثی کرده است. در نتیجه، مدل پیش‌بینی برتر گرادیان بوستینگ با اتکا به ویژگی‌های مهندسی‌شده، الگوهای غیرخطی ارقام را با بیشترین سطح دقت استخراج نموده است.



شکل ۵- مقایسه معماری عملکرد و خطای مدل‌های مختلف پیش‌بینی عملکرد نیشکر

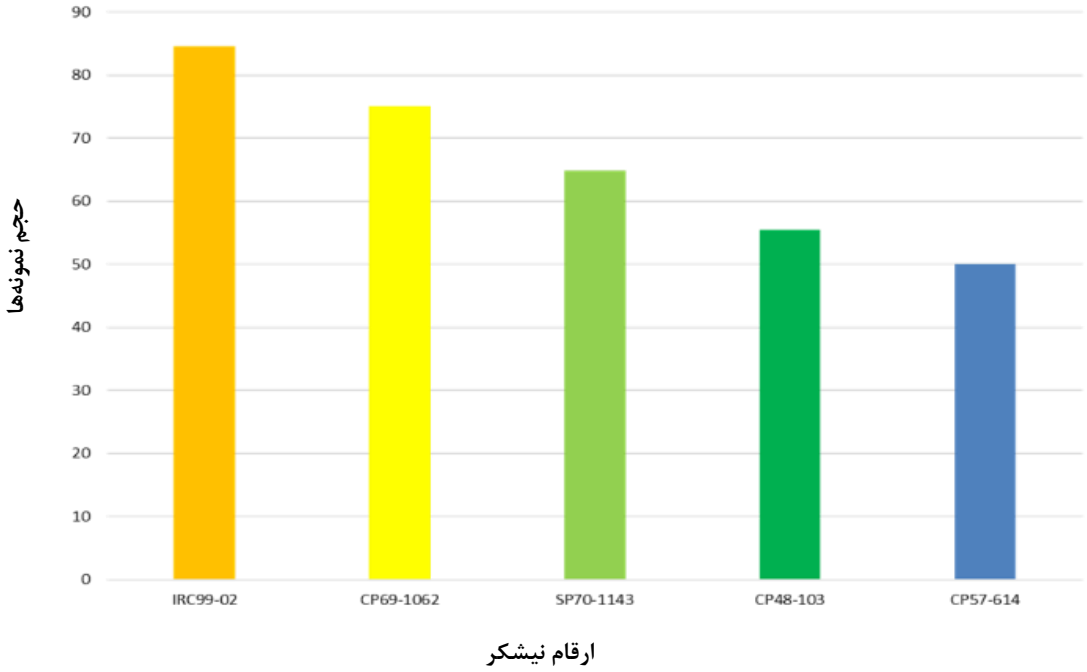
در نهایت، هرچند تحلیل تنوع عملکردی ارقام بر مبنای بازده ناخالص نهایی مزارع استوار بوده است، اما از منظر مدیریت پویای زراعی، پتانسیل بالایی برای توسعه این مدل‌ها در طول فصل رشد وجود دارد. در این راستا، با تکیه بر پایش سری زمانی طیفی (نظیر شاخص‌های LAI و EVI) و تلفیق آن با الگوریتم‌های موازنه انرژی سنجش از دوری جهت محاسبه تبخیر_تعرق واقعی (ETa)، می‌توان کارایی مصرف آب (WUE) ارقام مختلف را پیش از رسیدن به فاز برداشت نهایی به‌صورت برخط تخمین زد. این رویکرد آینده‌نگرانه، مبنای کلیدی برای نظام‌های پشتیبان تصمیم‌گیری در آبیاری هوشمند اراضی نیشکر خوزستان خواهد بود.

جدول ۵- آمار توصیفی بازده ناخالص نهایی ارقام نیشکر و حجم نمونه‌های مورد مطالعه

رقم	میانگین (تن/هکتار)	انحراف معیار	تعداد نمونه
IRC99-02	۸۴/۵۵	۳۱/۱۵	۱۱۲
CP69-1062	۷۵/۱۱	۱۹/۹۴	۱۱۵۰
SP70-1143	۶۴/۹۳	۱۹/۲۸	۴۲۸
CP48-103	۵۵/۴۴	۱۸/۳۸	۴۲۶
CP57-614	۵۰/۰۸	۱۴/۷۶	۳۰۱

تحلیل تفکیکی عملکرد مدل پیش‌بینی در خوشه‌های مدیریتی

به‌منظور ارزیابی پایداری ساختاری و سنجش میزان تعمیم‌پذیری الگوریتم توسعه‌یافته در شرایط متنوع زراعی، پتانسیل پیش‌بینی مدل برتر گرادیان بوستینگ به‌صورت تفکیک‌شده در هر یک از خوشه‌های مدیریتی مورد واکاوی قرار گرفت (جدول ۶ و شکل ۷). همان‌طور که ارقام استخراج‌شده گواهی می‌دهند، مدل الگوریتمی در آبر خوشه مزارع پربازده (شامل خوشه‌های ۲و ۴) به ضریب تبیین چشمگیر ۰/۹۶۷۰ با ریشه میانگین مربعات خطای کنترل‌شده ۰/۰۶۱۲ دست یافته است. این شاخص در آبر خوشه مزارع کم‌بازده و میان‌بازده (شامل خوشه‌های ۱و ۳) به اوج پایداری خود یعنی ۰/۹۹۳۶ می‌رسد. ثبت چنین سنجه‌های دقیقی در دو خوشه اصلی و فرکانس بالا در منطقه مورد مطالعه، تایید می‌کند که فرایند پیش‌پردازش و دسته‌بندی مزارع بر اساس بردارهای همگنِ شوری، بیوماس ماهواره‌ای و الگوی مصرف نهاده‌ها، به طور کاملاً موثری توانسته است اثرات مخرب نویزهای موضعی و واریانس خطای درون‌گروهی را مهار سازد.



شکل ۶- آمار توصیفی بازده ناخالص نهایی ارقام نیشکر و حجم نمونه‌های مورد مطالعه

از منظر تفسیر زراعی-محاسباتی، تفاوت اندک ضریب تبیین بین این دو محدوده، ریشه در رفتار بیوفیزیکی و فیزیولوژیک ارقام نیشکر دارد. در ابرخوشه مزارع کم‌بازده (خوشه‌های ۱ و ۳)، به دلیل غالب بودن ارقامی با پایداری رفتاری یکنواخت و انحراف معیار پایین‌تر در سطح منطقه (مانند رقم CP57-614)، پراکندگی داده‌ها پیرامون خط بهینه کمتر بوده و در نتیجه الگوریتم گرادیان بوستینگ با سهولت بیشتری الگوهای حاکم را استخراج کرده و ضریب تبیین بالاتری (۰/۹۹۳۶) برجای گذاشته است. در نقطه مقابل، در ابرخوشه مزارع پربازده (خوشه‌های ۲ و ۴)، حضور ژنوتیپ‌های حساسی نظیر IRC99-02 که دارای انحراف معیار بالایی بوده و واکنش‌های نوسانی و شدیدی به فاکتورهای محیطی و مدیریتی نشان می‌دهند، منجر به افزایش پیچیدگی الگوهای غیرخطی شده است، بااین‌حال، مدل گرادیان بوستینگ با بهره‌گیری از ساختار درختی ترتیبی و بهینه‌سازی توابع زیان، توانسته است این نوسانات را به‌خوبی مهار کرده و دقت فوق‌العاده بالایی (۰/۹۶۷۰) را ثبت نماید.

این پایداری تفکیکی، یک گام روبه‌جلو در مدیریت هوشمند کشاورزی به شمار می‌رود، چرا که الگوهای رگرسیون کلاسیک معمولاً در مواجهه با ناهمگنی‌های فاحش بیوفیزیکی دچار پدیده بیش برازش شده و در خوشه‌های بحرانی با افت شدید کارایی مواجه می‌شوند، درحالی‌که متدولوژی پیشنهادی این تحقیق، پتانسیل عملیاتی خود را در تمامی سطوح واریانس اراضی اثبات نموده است.

جدول ۶- عملکرد مدل گرادیان بوستینگ در هر خوشه

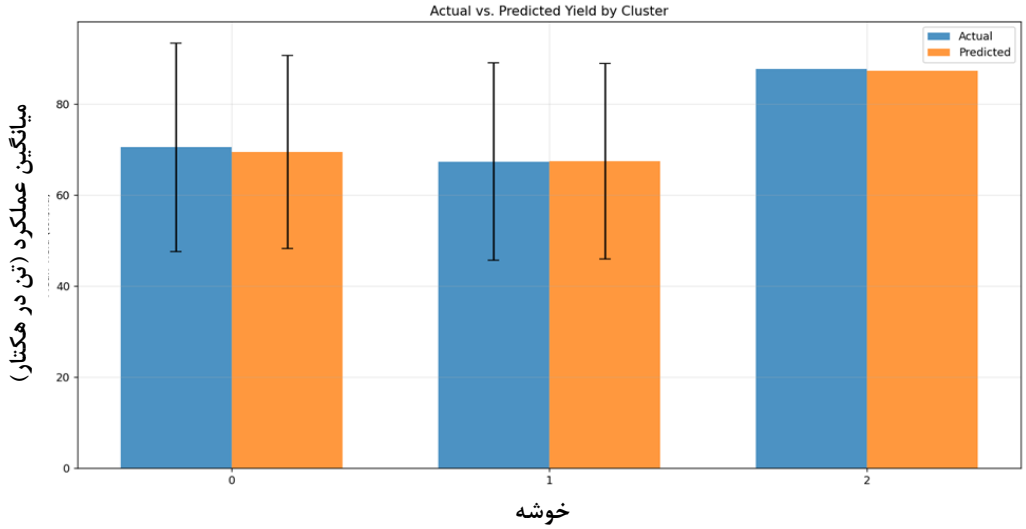
خوشه	ضریب تبیین	ریشه میانگین مربعات خطا
ابر خوشه مزارع پربازده (خوشه‌های ۲ و ۴)	۰/۹۶۷۰	۰/۰۶۱۲
ابر خوشه مزارع کم‌بازده (خوشه‌های ۱ و ۳)	۰/۹۹۳۶	۱/۷۳۶۳

مقایسه با مطالعات مشابه

تحلیل دستاوردهای این تحقیق در کنار چارچوب‌های پژوهشی گذشته، روند تکاملی و پایداری متدولوژی‌های مبتنی بر هوش مصنوعی را در بخش کشاورزی به‌خوبی نشان می‌دهد. در همین راستا، (Kaydani *et al.*, ۲۰۲۵) پتانسیل بالای تلفیق تصاویر ماهواره‌ای لندست و سنتینل را در برآورد دقیق عملکرد اراضی نیشکر هفت‌تپه اثبات نمودند. در پژوهش حاضر نیز، هم‌راستا با این یافته‌ها، از قابلیت‌های چندزمانی داده‌های سنجش از دوری استفاده شد، با این تفاوت که جهت کنترل چالش اشباع شاخص‌های طیفی در کانوپی‌های بسته نیشکر، متغیرهای مدیریتی زمینی نیز به‌عنوان مکمل به مدل تزریق گردیدند. همچنین، نتایج این تحقیق با یافته‌های (Mariadass

(*et al.*, ۲۰۲۲) که کارایی الگوریتم‌های درخت مبنا نظیر XGBoost را در شرایط اراضی ناهمگن تایید کردند، همخوانی کاملی دارد. مدل گرادیان بوستینگ به‌کاررفته در این پژوهش توانسته است با اتکا به ساختار درختی ترتیبی خود، روابط غیرخطی حاکم بر عملکرد را در خوشه‌های مختلف مدیریتی بازخوانی کند. این تقارن نتایج نشان می‌دهد که تلفیق رویکردهای یادگیری ماشین با سنجه‌های میدانی، مسیر پایداری را برای ارتقای پایش هوشمند در شرایط پیچیده زراعی فراهم می‌آورد.

پیرامون ساختار متدولوژیک توسعه‌یافته، لازم به ذکر است که ادغام شاخص‌های تعاملی مدیریت زراعی (نظیر کارایی مصرف آب) و اعمال خوشه‌بندی بر کل اراضی، علی‌رغم دستیابی به‌دقت پیش‌بینی ممتاز و بی‌سابقه، پتانسیل ایجاد نوعی بیش برازش موضعی یا نشت اطلاعات را در مدل‌های پیش‌بینی به همراه دارد. اگرچه پایداری و حجم بالای داده‌های زراعی خوزستان اثر این سوگیری را در این پژوهش تعدیل کرده است، اما پیشنهاد می‌شود در پژوهش‌های آتی، فرایندهای پیش‌پردازش، خوشه‌بندی K-means و اعتبارسنجی مدل‌ها مستقلاً و صرفاً بر روی داده‌های آموزش واسنجی شوند تا نرخ تعمیم‌پذیری مدل در مواجهه با مزارع کاملاً جدید ارتقا یابد.



شکل ۷- عملکرد مدل گرادیان بوستینگ در خوشه‌های مختلف

نتیجه‌گیری

در این پژوهش، یک چارچوب جامع، یکپارچه و بهینه‌سازی‌شده برای پیش‌بینی دقیق عملکرد ناخالص نیشکر بر پایه ادغام کلان‌داده‌های مدیریت زراعی و سنجه‌های ماهواره‌ای سنتینل ۲ توسعه یافت. دستاوردهای این تحقیق اثبات کرد که با ترکیب روش‌های خوشه‌بندی پیشرفته و معماری‌های یادگیری هوشمند جمعی، می‌توان به تخمین‌هایی فوق‌العاده پایدار و قابل‌تعمیم از وضعیت بیوماس محصول دست‌یافت. در متدولوژی پیشنهادی، ابتدا اراضی ناهمگن منطقه بر مبنای متغیرهای شوری، فیزیک خاک و شاخص‌های گیاهی با استفاده از الگوریتم K-means به چهار خوشه مدیریتی همگن تفکیک شدند؛ سپس مدل‌های رگرسیونی جنگل تصادفی و گرادیان بوستینگ به‌عنوان الگوریتم‌های ترکیبی گروهی بر روی فضای ویژگی‌های غنی‌شده آموزش دیدند.

دستاورد محوری این تحقیق، واسنجی و دستیابی به ضریب تبیین ممتاز ۰/۹۹۲۴ در مدل گرادیان بوستینگ بود که نقطه عطفی در مدل‌سازی هوشمند نیشکر به شمار می‌رود. نوآوری کلیدی پژوهش که ضامن پایداری این نرخ تعمیم‌پذیری شد، معرفی و اعتبارسنجی شاخص‌های ترکیبی کارآمد زراعی بود. تحلیل اهمیت ویژگی‌ها آشکار ساخت که سنجه‌های مهندسی‌شده کارایی مصرف آب و کود، سهم غالبی را در کاهش خطای مدل و تبیین تغییرات عملکرد اراضی ایفا می‌کنند. این یافته نه‌تنها قدرت توضیحی بالای روابط غیرخطی در الگوریتم‌های ترکیبی گروهیِ درخت مبنا را تایید می‌کند، بلکه بر ضرورت تغییر الگوهای تصمیم‌گیری از شاخص‌های گیاهی نوری محض (نظیر NDVI سنتی) به

سمت سنجه‌های تعاملی مدیریت مصرف آب و تغذیه در مزارع مدرن نیشکر تأکید دارد.

سپاسگزاری

نویسندگان از معاونت پژوهشی دانشگاه شهید چمران اهواز بابت تأمین هزینه‌های این پژوهش سپاسگزاری می‌نمایند.

مشارکت نویسندگان

نحوه و میزان مشارکت نویسندگان در انجام این پژوهش به‌صورت زیر است:

-نویسنده اول: جمع‌آوری و تحلیل داده‌ها و نگارش اولیه

-نویسنده دوم: طراحی مطالعه، بازبینی و ویرایش متن، ارائه نظرات تخصصی و بررسی‌های کارشناسی

دسترسی به داده‌ها

قابل‌انطباق نیست.

اصول اخلاقی

نویسندگان اصول اخلاقی را در انجام و انتشار این اثر علمی رعایت نموده‌اند و این موضع مورد تایید همه آنها است.

تضاد منافع نویسندگان

نویسندگان اعلام می‌کنند که هیچ‌گونه منافع مالی رقابتی یا روابط شخصی شناخته‌شده‌ای که ممکن است بر کار گزارش‌شده در این مقاله تأثیر گذاشته باشد، ندارند.

حمایت مالی

نویسندها حمایت مالی خاصی برای انجام این پژوهش دریافت نکرده‌اند.

منابع

Afsharnia, F & ,.Marzban, A. (2019). Application of integrated FMEA-ANP approach in prioritizing the risk of sugarcane transport delays to the factory. Journal of Agricultural Machinery, 9(2), 455-467. (In Persian). https://doi.org/10.22067/jam.v9i2.69447

Akbarian, S., Jamnani, M. R., Xu, C., Wang, W & ,. Lim, S. (2024). Plot level sugarcane yield estimation by machine learning on multi-spectral images :A case study of Bundaberg, Australia. Information Processing in Agriculture ,11(4),476-487 . https://doi.org/10/1016/j.inpa.2023.06.004

Barbosa, L. A. F., Pedronette, D. C. G & ,.Guilherme, I. R. (2023). A Meta-Feature Model for Exploiting Different Regressors to Estimate Sugarcane Crop Yiel d. IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium , https://doi.org/10.1109/IGARSS52108.2023.10283309

Canata, T. F., Wei, M. C. F., Maldaner, L. F & ,.Molin, J. P .(۲۰۲۱) .Sugarcane Yield Mapping Using High-Resolution Imagery Data and Machine Learning Technique. Remote Sens. 2021, 13, 232.

Bhatt, R., Hossain, A., Majumder, D., Chandra, M. S., Ghimire, R ,.Shahzad, M. F., Verma, K. K., Riar, A. S., Rajput, V. D & ,.Oliveira, M. W .(۲۰۲۴) . Prospects of artificial intelligence for the sustainability of sugarcane production in the modern era of climate change: An overview of related global findings. Journal of Agriculture and Food Research, 18, 101519 . https://doi.org/10.1016/j.jafr.2024.101519

Charoen-Ung, P & ,.Mittrapiyanuruk, P. (2018). Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning. International Conference on Computing and Information Technology , https://doi.org/10.1007/978-3-319-93692-5_4

Das, A., Kumar, M., Kushwaha, A., Dave, R., Dakhore, K. K., Chaudhari, K & ,.Bhattacharya, B. K. (2023). Machine learning model ensemble for predicting sugarcane yield through synergy of optical and SAR remote sensing .Remote Sensing Applications: Society and Environment, 30, 100962 . https://doi.org/10.1016/j.rsase.2023.100962

de França e Silva, N. R., Chaves, M. E. D., Luciano, A. C. d. S., Sanches ,I. D. A., de Almeida, C. M & ,. Adami, M. (2024). Sugarcane yield estimation using satellite remote sensing data in empirical or mechanistic modeling: A systematic review. Remote Sensing, 16(5), 863 . https://doi.org/10.3390/rs16050863

de Oliveira Maia, F. C., Bufon, V. B & ,.Leão, T. P. (2023). Vegetation indices as a tool for mapping sugarcane management zones. Precision Agriculture, 24 .۲۱۳-۲۳۴ ,(۱)https://doi.org/10.1007/s11119-022-09939-7.

Li, H., Wang, Z., Sun, L., Zhao, L., Zhao, Y., Li, X., Han, Y., Liang, S & ,.Chen, J. (2024). Parcel-Based Sugarcane Mapping Using Smoothed Sentinel-1 Time Series Data. Remote Sensing, 16(15), 2785 . https://doi.org/10.3390/rs16152785

Karami, P & ,.Esmaili-Rineh, S. (2020). Using Maximum Entropy (MaxEnt) modeling and K-mean clustering to analyze the underground habitats of Niphargus genus in Iran. Journal of Animal Research (Iranian Journal of Biology), 33, 310-323, (In Persian).

Kaydani, M., Hooshmand, A & ,.Hamzeh, S. (2025). Sugarcane yield estimation using Landsat and Sentinel satellite imagery (Case study: Haft Tappeh Sugarcane Agro-Industry). Iranian Journal of Irrigation and Drainage ,18(6), 1023-1033. (In Persian).

Corn Belt. Scientific reports, 11(1), 1606 . https://doi.org/10.1038/s41598-020-80820-1

Shendryk, Y., Davy, R & ,.Thorburn, P. (2021). Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. Field Crops Research, 260, 107984 . https://doi.org/10.1016/j.fcr.2020.107984

Seyed Jalali, S. M., Navidi, M. N & ,.Zeinadini Maymand, M. (2021) .Estimating sugarcane production potential with different models in the lands of southern Khuzestan province. Journal of Soil Research, 35(1), 59-74. (In Persian) https://doi.org/10.22092/ijsr.2021.351600.548

Singla, S. K., Garg, R. D & ,.Dubey, O. P. (2021). Ensemble machine learning methods for spatio-temporal data analysis of plant and ratoon sugarcane. Intelligent Data Analysis, 25(5), 1291-1322.

Soltani-Kazemi, M., Minaei, S., Shafi-Zadeh-Moghadam, M & ,.Mahdavian ,A. (2024). Comparison of four algorithms PLSR, RF, GRNN, and SVR for estimating sugarcane canopy moisture during the growth period using Sentinel-2 satellite imagery. Iranian Journal of Remote Sensing and GIS, 16(3), 47-68. (In Persian) . https://doi.org/10.48308/gisj.2023.103163

Taravat, A., Abebe, G., Gessesse, B & ,.Tadesse, T. (2024). Estimation of Sugarcane Yield Using Multi-Temporal Sentinel 2 Satellite Imagery and Random Forest Regression .http://dx.doi.org/10.5194/isprs-archives-XLVIII-4-W9-2024-357-2024

Van Klompenburg ,T., Kassahun, A & ,.Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and electronics in agriculture, 177, 105709 . https://doi.org/10.1016/j.compag.2020.105709

Zhang, S., Qin, H., Li, X., Zhang ,M., Yao, W., Lyu, X & ,.Jiang, H. (2025). Sensitive Multi-spectral Variable Screening Method and Yield Prediction Models for Sugarcane Based on Gray Relational Analysis and Correlation Analysis. Remote Sensing, 17(12), 2055 .https://doi.org/10.3390/rs17122055

Mariadass, D. A., Moung, E. G., Sufian, M. M & ,. Farzamnia, A. (2022 ,November). EXtreme gradient boosting (XGBoost) regressor and shapley additive explanation for crop yield prediction in agriculture. In 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE) 219-224 . https://doi.org/10/1109/ICCKE57176.2022.9960069

Mirzaei, M., Marofi, S., Solgi, E., Abbasi, M & ,. Karimi, R .(2021) .Non-destructive estimation of chlorophyll and nitrogen in grape leaves using ground-based hyperspectral data and support vector machine method. Journal of Plant Research (Iranian Journal of Biology), 34(1), 191-204. (In Persian). https://doi.org/20.1001.1.23832592.1400.34.1.12.9

Nakhaeinejad, M., Abbasi, M., ZareMehrjerdy, Y & ,. Asadi Zarch, A. (2022) .Greenhouse gas emission reduction model: an integrated approach of linear programming and system dynamics: The case of Iranian power plants .Research in Production and Operations Management, 13(1), 51-77. https://doi.org/10.22108/jpom.2022.129313.1382

Parashar, N., Johri, P., Khan, A. A., Gaur, N & ,.Kadry, S. (2024). An Integrated Analysis of Yield Prediction Models: A Comprehensive Review of Advancements and Challenges. Computers, Materials & Continua, 80 .http://dx.doi.org/10.32604/cmc.2024.050240

Poudyal, C., Costa, L. F., Sandhu, H., Ampatzidis, Y., Odero, D. C ,.Arbelo, O. C & ,.Cherry, R. H. (2022). Sugarcane yield prediction and genotype selection using unmanned aerial vehicle-based hyperspectral imaging and machine learning. Agronomy Journal, 114(4), 2320–2333 . https://doi.org/10.1016/j.fcr.2020.107984

Parchami-Araghi, F. & Sadeghi-Lari, A. (2021). Uncertainty analysis of distributed and sub-daily agro-hydrological modeling of a sugarcane cropping system under combined free/controlled subsurface drainage management. Iran Water Research Journal, 15(3), 37-49. (In Persian).

Shahhosseini, M., Hu, G., Huber, I & ,.Archontoulis, S. V .(2021) .Coupling machine learning and crop modeling improves crop yield prediction in the US