



Modeling Nonlinear Soil-Vegetation Relationships in the Zouji Shoush Atershed: A Machine Learning Approach with Uncertainty Quantification

Kohzad Heidary^{1*}, Mehri Dinarvand², Saba Peyrov³, Seyed Hossein Arami⁴

1. Corresponding Author, Assistant Professor, Soil Conservation and Watershed Management Research Department, Khuzestan Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Ahvaz, Iran, Email: Kohzad93@gmail.com
2. Associate Professor, Forests and Rangelands Research Department, Research Institute of Forests and Rangelands, Agricultural Research, Education and Extension Organization (AREEO), Ahvaz, Iran, Email: mehri.dinarvand@gmail.com
3. Assistant Professor, Forests and Rangelands Research Department, Khuzestan Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Ahvaz, Iran, Email: peyrov@saba@gmail.com
4. Assistant Professor, Forests and Rangelands Research Department, Khuzestan Agricultural and Natural Resources Research and Education Center, Agricultural Research Education and Extension Organization (AREEO), Ahvaz, Iran Email: arami1854@gmail.com & arami1854@areeo.ac.ir

Article Info

Article type:
Research Article

Article history:

Received 28 March 2026

Revised 01 May 2026

Accepted 13 June 2026

Available online 22 June 2026

Keywords:

Machine Learning,
Soil Characteristics,
Uncertainty Quantification,
Gradient Boosting,
Zojy Shush Watershed

ABSTRACT

Research Topic: Modeling Vegetation Cover Dynamics in Arid Regions Using Deep Learning and Multi-Source Data

Objective: This study aimed to model vegetation cover percentage based on soil characteristics in the Zojy Shush watershed, Khuzestan Province, Iran. Two topographic units (hill and valley) were considered. Machine learning algorithms were employed alongside interpretability and uncertainty quantification approaches.

Method: Following sampling from 120 points, nine soil characteristics (pH, phosphorus, potassium, organic carbon, EC, lime, clay, silt, and sand) and vegetation cover percentage were measured. Three algorithms—Random Forest, XGBoost, and Gradient Boosting—were employed with cross-validation. Feature importance was assessed using Gini, permutation, and SHAP indices, while uncertainty was evaluated via the bootstrap method.

Results: In the topographically-separated model for the Valley region, Gradient Boosting achieved a median R^2 of 0.73, outperforming XGBoost (0.68) and Random Forest (0.14). In contrast, for the combined dataset, XGBoost showed superior performance (median $R^2 = 0.50$). In contrast, the integrated model (combining both hills and valleys) yielded negative R^2 values due to topographic heterogeneity. Clay (0.245), organic matter (0.156), and phosphorus (0.130) were identified as the most important predictors. Gradient Boosting provided the best uncertainty coverage (60%), with epistemic uncertainty being dominant across the models. Topography-specific modeling (separate models for hills and valleys with performance around $R^2 \approx 0.78$) demonstrated a significant advantage over a unified model (which yielded negative R^2 values). Furthermore, phosphorus, pH, and silt were identified as primary limiting factors for vegetation cover in 75%, 50%, and 25% of the study area, respectively.

Conclusions: This study revealed that soil-vegetation relationships in arid regions are nonlinear and dependent on topographic heterogeneity. Phosphorus was identified as the primary limiting factor in calcareous soils. Despite a weak linear correlation, clay emerged as the most important predictor due to its threshold-based relationship with vegetation cover. The predominance of epistemic uncertainty suggests that enhancing the modeling framework is more effective than merely increasing measurement accuracy for improving predictions in such complex environments.

Cite this article: Heidary, K., Dinarvand, M., Peyrov, S., Arami, S. H. (2026). Modeling Nonlinear Soil-Vegetation Relationships in the Zouji Shoush Atershed: A Machine Learning Approach With Uncertainty Quantification. *Journal of Ecohydrology*, 13(2), 1316-1341.



<https://doi.org/10.22059/ije.2026.412192.1909>

© Author(s) retain the copyright.

Publisher: University of Tehran Press.

Extended Abstract

Introduction

Vegetation cover in arid and semi-arid regions is a critical indicator of ecosystem health, playing a vital role in soil erosion control, microclimate regulation, and combating desertification. The complex, non-linear, and threshold-dependent relationships between soil physicochemical properties and vegetation dynamics, however, pose significant challenges for accurate predictive modeling. While machine learning (ML) has advanced soil science, critical research gaps persist: a lack of simultaneous, bidirectional investigation of soil-vegetation systems; the common oversight of topographic heterogeneity in integrated modeling frameworks; and insufficient attention to model interpretability and the quantification of prediction uncertainty. This study aimed to address these gaps by modeling vegetation cover percentage based on soil properties across two distinct topographic units (hill and valley) in the Zouji Shoush watershed, Khuzestan province, Iran, using advanced ML algorithms coupled with interpretability and uncertainty analysis techniques.

Method

The study was conducted in the protected Zouji Shoush watershed (200 ha, warm semi-desert climate, ~300 mm annual precipitation). A systematic-random sampling strategy, stratified by topographic units, was employed to collect 120 composite soil samples (60 from hills, 60 from valleys). Nine key soil physicochemical properties pH, available phosphorus (P), exchangeable potassium (K), soil organic carbon (OC), electrical conductivity (EC), total neutralizing value (TNV, as a measure of calcium carbonate equivalent), and percentages of clay, silt, and sand were measured using standard laboratory methods. The target variable, vegetation canopy cover (%), was also recorded. Three tree-based ensemble algorithms Random Forest (RF), XGBoost (XGB), and Gradient Boosting (GB) were implemented with 5-fold cross-validation to predict vegetation cover from the soil properties. Feature importance was quantified using three complementary approaches: Gini impurity (model-based), permutation importance (model-agnostic), and SHAP (SHapley Additive exPlanations) values for local interpretability. Prediction uncertainty was comprehensively assessed through bootstrap resampling (50 iterations) by calculating prediction intervals (PIs), prediction interval coverage probability (PICP), and mean PI width. Total uncertainty was further decomposed into its epistemic (model knowledge deficiency) and aleatoric (inherent data variability) components. Modeling was performed separately for hill and valley subgroups as well as in an integrated manner to evaluate the impact of topographic heterogeneity.

Results

Pearson correlation analysis revealed strong intra-soil relationships, with the most negative correlations observed between sand-clay ($r = -0.89$) and between TNV with clay ($r = -0.92$) and silt ($r = -0.94$), indicating a strong association of carbonates with fine soil fractions. TNV showed the strongest negative linear correlation with vegetation ($r = -0.19$). A regional analysis of 10 representative zones identified phosphorus (limiting in 75% of zones), pH (50%), and silt (25%) as the primary local limiting factors for vegetation. In the topographically-separated model for the Valley region, Gradient Boosting demonstrated superior predictive performance (median $R^2 = 0.73$) compared to XGBoost (0.68) and Random Forest (0.14). However, in the combined dataset (All Data), XGBoost achieved the highest median R^2 (0.50), followed by Gradient Boosting (0.47) and Random Forest (0.37). Normalized feature importance from the GB model identified clay (importance weight: 0.245), organic carbon (0.156), phosphorus (0.130), TNV (0.128), and sand (0.113) as the top five consensus predictors. This ranking

highlights a key finding: despite its weak linear correlation, clay emerged as the most important predictor, suggesting its influence on vegetation is non-linear and likely threshold-dependent, an effect captured by the ML model but missed by simple correlation. The importance of P and TNV reaffirms their role as key chemical constraints in these calcareous soils. SHAP analysis consistently identified K, pH, and TNV as the most influential features across models, offering a nuanced, model-agnostic perspective on variable importance. Bootstrap analysis indicated variability in model performance across resampled datasets, with occasional negative R^2 values observed in some iterations. These results suggest sensitivity of the integrated models to environmental heterogeneity and sample composition. Therefore, median R^2 values were used as the primary metric for model comparison. In contrast, GB demonstrated lower mean performance ($R^2 = 0.016$) but significantly higher stability ($SD = 0.152$) and provided the best prediction interval coverage (PICP = 60%, mean PI width = 18.7%), suggesting more reliable uncertainty estimates. Critically, the decomposition of uncertainty revealed that epistemic uncertainty was dominant (>50% of total uncertainty) across all models. This indicates that the primary source of prediction error stems from model limitations, such as missing process-based variables or insufficiently captured complex interactions, rather than from random measurement error. Furthermore, subgroup modeling based on topographic units yielded substantially higher predictive performance ($R^2 \approx 0.78$) compared to the integrated model that pooled data from both hills and valleys, which exhibited poor performance (negative R^2 values). This confirms the "heterogeneity penalty" concept, demonstrating that amalgamating data from inherently different environmental contexts can degrade model accuracy.

Conclusions

This study demonstrates that soil-vegetation relationships in arid regions are inherently non-linear, threshold-dependent, and contingent upon topographic heterogeneity. Gradient Boosting emerged as the superior algorithm due to its capacity for modeling complex non-linear interactions and providing more reliable uncertainty estimates. Phosphorus was identified as the primary limiting factor in these calcareous soils, emphasizing the need for targeted phosphorus management strategies in restoration efforts. The predominance of epistemic uncertainty suggests that future research should prioritize improving modeling paradigms by incorporating process-based variables (e.g., soil moisture dynamics, microbial activity) over merely increasing measurement precision. The clear superiority of topographically-differentiated models confirms that accounting for landscape heterogeneity is not optional but essential for accurate predictive modeling in complex terrains. This research provides a comprehensive and novel methodological framework that integrates advanced ML, multi-faceted interpretability, and rigorous uncertainty quantification, offering a valuable template for analogous studies in arid and semi-arid regions globally.

Acknowledgements

We hereby express our gratitude for the support provided by the Khuzestan Agricultural and Natural Resources Research and Education Center for conducting this research project.

Conflict of Interest

The authors of this article declare that they have no conflict of interest regarding the writing and publication of the materials and results of this research.

Data Availability

The data and results used in this research will be made available upon correspondence with the corresponding author.

Author Contributions

First Author: Formal analysis and statistical review of data, preparation and drawing of graphs and visual data.

Second Author: Conceptualization, formulation of methodology.

Third Author: Editing and final review of the article.

Fourth Author: Writing the initial draft of the article.



مدل سازی روابط غیر خطی خاک - گیاه در حوضه آبخیز زوجی شوش با رویکرد یادگیری ماشین و تحلیل عدم قطعیت

کهزاد حیدری^{۱*}، مهرداد دیناروند^۲، صبا پیرو^۳، سید حسین آرامی^۴

۱. (نویسنده مسئول)، استادیار پژوهشی بخش تحقیقات حفاظت خاک و آبخیزداری، مرکز تحقیقات، آموزش کشاورزی و منابع طبیعی خوزستان، سازمان تحقیقات، آموزش و ترویج کشاورزی، اهواز، ایران. پست الکترونیکی: kohzad93@gmail.com و k.heidary@areeo.ac.ir
۲. دانشیار پژوهشی، بخش تحقیقات جنگل ها و مراتع، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان خوزستان، سازمان تحقیقات، آموزش و ترویج کشاورزی، اهواز، ایران. mehri.dinarvand@gmail.com
۳. استادیار پژوهشی، بخش تحقیقات جنگل ها و مراتع، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان خوزستان، سازمان تات، peyrowsaba@gmail.com
۴. استادیار پژوهشی، بخش تحقیقات جنگل ها و مراتع، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان خوزستان، سازمان تحقیقات، آموزش و ترویج کشاورزی، اهواز، ایران. arami1854@gmail.com و arami1854@areeo.ac.ir

اطلاعات مقاله	چکیده
نوع مقاله: پژوهشی	موضوع: مدل سازی پویایی پوشش گیاهی در مناطق خشک با استفاده از یادگیری عمیق و داده های چندمنبعی
تاریخ دریافت: ۱۴۰۵/۰۱/۰۸	هدف: این مطالعه با هدف مدل سازی درصد پوشش گیاهی بر اساس ویژگی های خاک در دو واحد توپوگرافی تپه و دره در حوضه آبخیز زوجی شوش استان خوزستان با استفاده از الگوریتم های یادگیری ماشین و رویکردهای تفسیرپذیری و عدم قطعیت انجام شد.
تاریخ بازنگری: ۱۴۰۵/۰۲/۱۱	روش پژوهش: پس از نمونه برداری از ۱۲۰ نقطه، ۹ ویژگی خاک (اسیدیته، فسفر، پتاسیم، کربن آلی، EC، آهک، رس، سیلت و شن) و درصد پوشش گیاهی اندازه گیری شد. سه الگوریتم جنگل تصادفی، XGBoost و گرادیان بوستینگ با اعتبارسنجی متقابل به کار رفتند. اهمیت ویژگی ها با شاخص های جینی، جایگشتی و SHAP و عدم قطعیت با روش بوت استرپ ارزیابی شد.
تاریخ پذیرش: ۱۴۰۵/۰۳/۲۳	یافته ها: نتایج نشان داد در مدل تفکیکی برای منطقه دره، گرادیان بوستینگ با میانه R^2 برابر ۰/۷۳ عملکرد بهتری نسبت به XGBoost (۰/۶۸) و جنگل تصادفی (۰/۱۴) داشت. در مقابل، مدل یکپارچه (ترکیب تپه و دره) به دلیل ناهمگونی توپوگرافی، مقادیر R^2 منفی نشان داد. رس (۰/۲۴۵)، ماده آلی (۰/۱۵۶) و فسفر (۰/۱۳۰) مهم ترین پیش بین کننده ها بودند. گرادیان بوستینگ بهترین پوشش عدم قطعیت (۶۰ درصد) را ارائه داد و عدم قطعیت معرفتی بر مدل ها غالب بود. مدل سازی تفکیکی بر اساس واحدهای توپوگرافی (عملکرد حدود ۰/۷۸) نسبت به مدل یکپارچه (با عملکرد منفی) برتری چشمگیری داشت. همچنین فسفر، pH و سیلت به ترتیب در ۷۵ درصد، ۵۰ درصد و ۲۵ درصد مناطق به عنوان عوامل محدودکننده اولیه پوشش گیاهی شناسایی شدند.
کلیدواژه ها:	نتیجه گیری: این مطالعه نشان داد روابط خاک - گیاه در مناطق خشک غیرخطی و وابسته به ناهمگونی توپوگرافی است. فسفر عامل محدودکننده اصلی در خاک های آهکی شناسایی شد و علی رغم همبستگی خطی ضعیف، رس به دلیل رابطه آسانه ای با پوشش گیاهی به مهم ترین پیش بین کننده تبدیل شد. غلبه عدم قطعیت معرفتی نشان می دهد بهبود مدل سازی مؤثرتر از افزایش دقت اندازه گیری است.
یادگیری ماشین، ویژگی های خاک، عدم قطعیت، گرادیان بوستینگ	
حوضه آبخیز زوجی شوش.	

استاد: حیدری، کهزاد؛ دیناروند، مهرداد؛ پیرو، صبا؛ آرامی، سید حسین. (۱۴۰۵). مدل سازی روابط غیر خطی خاک-گیاه در حوضه آبخیز زوجی شوش با رویکرد

یادگیری ماشین و تحلیل عدم قطعیت. مجله اکوهیدرولوژی، ۱۳(۲)، ۱۳۱۶-۱۳۴۱.

<https://doi.org/10.22059/ije.2026.412192.1909>

ناشر: انتشارات دانشگاه تهران.



© نویسندگان: کهزاد حیدری، مهرداد دیناروند، سید حسین آرامی

مقدمه

خاک به عنوان بستر حیاتی اکوسیستم های خشکی، نقشی محوری در تأمین امنیت غذایی، تنظیم چرخه های زیست زمین شیمیایی و حفظ تنوع زیستی ایفا می کند. برآوردها نشان می دهد تقریباً ۹۸/۸ درصد از کالری مصرفی انسان از خاک منشأ می گیرد (کوپیتک^۱ و همکاران، ۲۰۱۹). با این حال، فشار فزاینده ناشی از تغییر اقلیم، تشدید کشاورزی و تخریب اراضی، سلامت خاک ها را با تهدید جدی مواجه ساخته است. مطالعات اقتصادی نشان می دهد هر دلار سرمایه گذاری در کاهش تخریب خاک، بازده اقتصادی ۷ تا ۳۰ دلاری به همراه دارد و این سرمایه گذاری می تواند ارزش تولیدات کشاورزی جهانی را تا ۲/۴ تریلیون دلار در سال افزایش دهد (فائو^۲، ۲۰۲۱). پوشش گیاهی در مناطق خشک و نیمه خشک (که حدود ۴۱ درصد از سطح زمین را می پوشانند) نه تنها شاخص کلیدی بهره وری اکوسیستم است، بلکه نقش حفاظتی حیاتی در کنترل فرسایش، کاهش تبخیر و مقابله با بیابان زایی ایفا می کند (یو^۳ و همکاران، ۲۰۲۱). با این وجود، روابط پیچیده و غیرخطی میان ویژگی های خاک و پوشش گیاهی در این مناطق، مدل سازی دقیق این تعاملات را با چالش مواجه ساخته است. این روابط اغلب آستانه ای، تعاملی و به عمق خاک و شرایط توپوگرافی وابسته هستند. در نتیجه، مدل سازی دقیق نیازمند رویکردهای تحلیلی پیشرفته ای است که قادر به شناسایی الگوهای غیرخطی باشند.

چارچوب مفهومی CLORPT که توسط جنی^۴ (۱۹۴۱) پایه گذاری شد، نخستین تلاش سیستماتیک برای تبیین نقش عوامل خاک ساز در تشکیل و تکامل خاک بود. این چارچوب بعدها توسط مک براتنی^۵ و همکاران (۲۰۰۳) به مدل SCORPAN تکامل یافت که با گنجانیدن مؤلفه های مکانی و طیفی، بستر نظری نقشه برداری رقومی خاک را فراهم آورد (میناسنی و مک براتنی^۶، ۲۰۱۶؛ وادو^۷ و همکاران، ۲۰۲۰). در دو دهه اخیر، همگرایی پیشرفت ها در سنجش از دور، افزایش قدرت محاسباتی و توسعه الگوریتم های یادگیری ماشین انقلابی در نقشه برداری و مدل سازی خاک پدید آورده است (وادو و همکاران، ۲۰۲۰). به خلاف روش های کلاسیک زمین آمار که بر مفروضات ایستایی و خطی بودن استوارند، الگوریتم های یادگیری ماشین قادر به شناسایی روابط غیرخطی و پیچیده، تشخیص تعاملات میان متغیرها و استخراج الگوهای چندبعدی از داده های حجیم و ناهمگن هستند (وادو و همکاران، ۲۰۲۰). مرور سیستماتیک توسط پاداریان^۸ و همکاران (۲۰۲۰) نشان می دهد الگوریتم های مبتنی بر درخت تصمیم، به ویژه جنگل تصادفی^۹ و روش های تقویتی مانند Gradient Boosting، پرمصرف ترین و موفق ترین روش ها در مدل سازی ویژگی های خاک در سال های اخیر بوده اند.

پیشینه پژوهش

مطالعات متعددی به کارگیری یادگیری ماشین را در پیش بینی ویژگی های خاک و پوشش گیاهی بررسی کرده اند. بالتسن وایلر^{۱۰} و همکاران (۲۰۲۱) با استفاده از مدل Random Forest به نقشه برداری ۱۳ ویژگی فیزیکی - شیمیایی خاک در جنگل های سوئیس پرداختند و عملکرد برتر این الگوریتم را در مقایسه با روش های سنتی نشان دادند. وینتسلر^{۱۱} و همکاران (۲۰۲۴) با تلفیق داده های سنجش از دور و یادگیری ماشین، دقت پیش بینی ویژگی های حاصلخیزی خاک از جمله pH، فسفر و پتاسیم را به طور معناداری

1 Kopittk

2 FAO

3 Yu

4 Jenny

5 McBratney

6 Minasny & McBratney

7 Wadoux

8 Padian

9 Random Forest

10 Baltensweiler

11 Winzeler

بهبود بخشیدند. فورکوئر^۱ و همکاران (۲۰۱۷) در مطالعه خود روی پیش‌بینی ویژگی‌های خاک با داده‌های سنجش از دور، ظرفیت بالای الگوریتم‌های یادگیری ماشین (مانند جنگل تصادفی) را در تخمین درصد شن، سیلت و رس در مقایسه با روش‌های رگرسیون خطی تأیید کردند. در زمینه روابط خاک - گیاه و پهنه‌بندی ویژگی‌های خاک در مناطق نیمه‌خشک، زراعت‌پیشه^۲ و همکاران (۲۰۱۹) با استفاده از روش‌های یادگیری ماشین نشان دادند متغیرهای محیطی و ویژگی‌های خاک (مانند بافت و کربن آلی) همبستگی بالایی با توزیع فضایی پوشش گیاهی و سایر پارامترها دارند. مطالعه یانگ^۳ و همکاران (۲۰۱۶) با به‌کارگیری مدل‌های درختی تقویتی^۴ در حوضه‌های آبخیز کوهستانی، نقش غالب توپوگرافی و پوشش گیاهی را در توزیع مکانی عمق خاک آشکار ساختند و با تحلیل اهمیت متغیرها، تأثیر این عوامل را در ارتفاعات مختلف مورد ارزیابی قرار دادند. ژانگ^۵ و همکاران (۲۰۲۱) با تلفیق داده‌های ماهواره‌ای چندمنبعی و الگوریتم‌های یادگیری ماشین، قابلیت پیش‌بینی عملکرد ذرت را بهبود بخشیدند و با تحلیل اهمیت متغیرها، نقش تعیین‌کننده ویژگی‌های محیطی و شرایط خاک را در توزیع مکانی عملکرد و پاسخ گیاه به تنش‌های محیطی نشان دادند. شاه‌حسینی^۶ و همکاران (۲۰۲۰) با توسعه و ارزیابی مدل‌های یادگیری ماشین پیشرفته از جمله XGBoost برای پیش‌بینی عملکرد ذرت، نشان دادند ویژگی‌های خاک (نظیر بافت و ظرفیت نگهداری آب) در کنار متغیرهای مدیریتی و اقلیمی، نقش تعیین‌کننده‌ای در کاهش خطای مدل و توجیه تغییرات مکانی عملکرد دارند. همچنین، پژوهش‌های انجام‌شده در ایران نیز نشان‌دهنده ظرفیت بالای مدل‌سازی رقومی خاک با الگوریتم‌های یادگیری ماشین است. برای مثال، موسوی^۷ و همکاران (۲۰۲۰) در مطالعه‌ای درباره پیش‌بینی مکانی ویژگی‌های خاک سطحی شامل کربن آلی، کربنات کلسیم معادل، رس، سیلت و شن در منطقه‌ای در کردستان، مدل Random Forest و روش‌های درون‌یابی با داده‌های مدل رقومی ارتفاع و تصاویر ماهواره‌ای را به کار گرفتند و نتایج معتبری ارائه دادند.

علی‌رغم پیشرفت‌های قابل توجه در کاربرد یادگیری ماشین در علوم خاک و بوم‌شناسی، مرور نظام‌مند ادبیات پژوهشی سه شکاف اساسی را آشکار می‌سازد که این پژوهش با هدف پوشش آن‌ها انجام شده است. نخست، عدم بررسی هم‌زمان دوسویه خاک - گیاه: اکثر مطالعات پیشین یا بر پیش‌بینی ویژگی‌های خاک متمرکز بوده‌اند (موسوی و همکاران، ۲۰۲۰؛ بالتنس‌وایلر و همکاران، ۲۰۲۱؛ وینتسلر و همکاران، ۲۰۲۴) یا بر تخمین زیست‌توده و پوشش گیاهی، و به بررسی هم‌زمان و دوسویه این دو مؤلفه نپرداخته‌اند. این شکاف زمانی پررنگ‌تر می‌شود که ویژگی‌های خاک هم به عنوان پیش‌بین و هم به عنوان پاسخ در سیستم خاک - گیاه عمل می‌کنند. دوم، نادیده گرفتن ناهمگونی توپوگرافی در مدل‌سازی یکپارچه: بسیاری از مطالعات، واحدهای توپوگرافی متمایز (تپه، دره، شیب) را بدون تفکیک در یک مدل یکپارچه ترکیب می‌کنند. مایر^۸ و همکاران (۲۰۱۹) مفهوم «عارضه ناهمگونی» را معرفی کردند و نشان دادند مدل‌سازی یکپارچه مناطق ناهمگن، علی‌رغم افزایش حجم نمونه، به کاهش دقت پیش‌بینی منجر می‌شود؛ زیرا مدل سعی در یادگیری «میانگین» روابط دارد که برای هیچ‌یک از زیرگروه‌ها به طور بهینه صدق نمی‌کند. مطالعات در اکوسیستم‌های پیچیده کوهستانی نظیر فلات تبت و کوهستان‌های مرتفع نشان داده‌اند تفکیک واحدهای همگن پدوژنتیکی برای غلبه بر مشکل ناهمگونی محیطی ضروری است (یانگ، ۲۰۱۶) سوم، به رغم پیشرفت‌های اخیر، به کمی‌سازی عدم قطعیت و تفسیرپذیری مدل‌ها در مدل‌سازی خاک - گیاه در مقایسه با دقت پیش‌بینی صرف، توجه نسبتاً کمتری شده است: ارزیابی عدم قطعیت پیش‌بینی‌ها و تفکیک منابع خطا به مؤلفه‌های ایستیمیک ناشی از محدودیت دانش مدل و آلتوری ناشی از تغییرپذیری

1 Forkuor

2 Zeraatpisheh

3 Yang

4 BRT

5 Zhang

6 Shahhosseini

7 Mousavi

8 Meyer

ذاتی داده‌ها در اغلب مطالعات خاک - گیاه نادیده گرفته شده است (وادو^۱ و همکاران، ۲۰۲۰؛ دیتسه^۲، ۲۰۱۷). این در حالی است که برای کاربردهای مدیریتی و تصمیم‌سازی، صرف داشتن پیش‌بینی نقطه‌ای کافی نیست و کمی‌سازی عدم قطعیت برای ارزیابی ریسک و قابلیت اطمینان ضروری است. افزون بر این، استفاده از روش‌های تفسیرپذیری مانند SHAP که امکان شناسایی روابط غیرخطی، آستانه‌ای و تعاملی را در سطح نمونه‌های فردی فراهم می‌کند، تا همین اواخر محدود به مطالعات معدودی بوده است (لوندبرگ و لی^۳، ۲۰۱۷).

با توجه به شکاف‌های یادشده، مطالعه‌ای که به طور هم‌زمان ویژگی‌های فیزیکی و شیمیایی خاک را در واحدهای توپوگرافی متمایز (تپه و دره)، و درصد پوشش گیاهی را در چارچوبی یکپارچه با بهره‌گیری از الگوریتم‌های پیشرفته یادگیری ماشین، تحلیل اهمیت چندروشی و کمی‌سازی عدم قطعیت بررسی کند، یک ضرورت انکارناپذیر پژوهشی است. این پژوهش برای نخستین بار در آبخیز معرف زوجی شوش (استان خوزستان) با رویکردی جامع و نوآورانه به این شکاف‌ها پاسخ می‌دهد. این مطالعه با اهداف زیر طراحی و اجرا شده است: مدل‌سازی رابطه غیرخطی؛ مدل‌سازی روابط غیرخطی میان ۹ ویژگی فیزیکی و شیمیایی خاک و درصد تاج پوشش گیاهی با استفاده از سه الگوریتم یادگیری ماشین مبتنی بر درخت تصمیم (Random Forest، XGBoost و Gradient Boosting)؛ تحلیل اهمیت متغیرها؛ شناسایی و رتبه‌بندی مهم‌ترین ویژگی‌های خاک مؤثر بر پوشش گیاهی با سه رویکرد مکمل مستقل (اهمیت درون مدلی، اهمیت جایگشتی و مقادیر SHAP)؛ اثر توپوگرافی؛ کمی‌سازی واحد توپوگرافی (تپه و دره) بر روابط خاک - گیاه و ارزیابی عملکرد مدل‌های تفکیک‌شده در برابر مدل یکپارچه؛ کمی‌سازی عدم قطعیت؛ کمی‌سازی عدم قطعیت پیش‌بینی‌ها با روش بوت‌استرپ و تفکیک منابع خطا به مؤلفه‌های اپیستمیک و آلیاتوریک.

نتایج این پژوهش می‌تواند به بهبود مدیریت منابع طبیعی و تدوین راهبردهای هدفمند احیای پوشش گیاهی در خاک‌های آهکی با محدودیت فسفر را تدوین کند؛ چارچوب روش‌شناختی برای مدل‌سازی خاک - گیاه در مناطق ناهمگن توپوگرافی فراهم آورد و به عنوان الگویی برای پژوهش‌های مشابه در سایر مناطق خشک و نیمه‌خشک کشور و منطقه غرب آسیا مورد استفاده قرار گیرد.

روش‌شناسی پژوهش

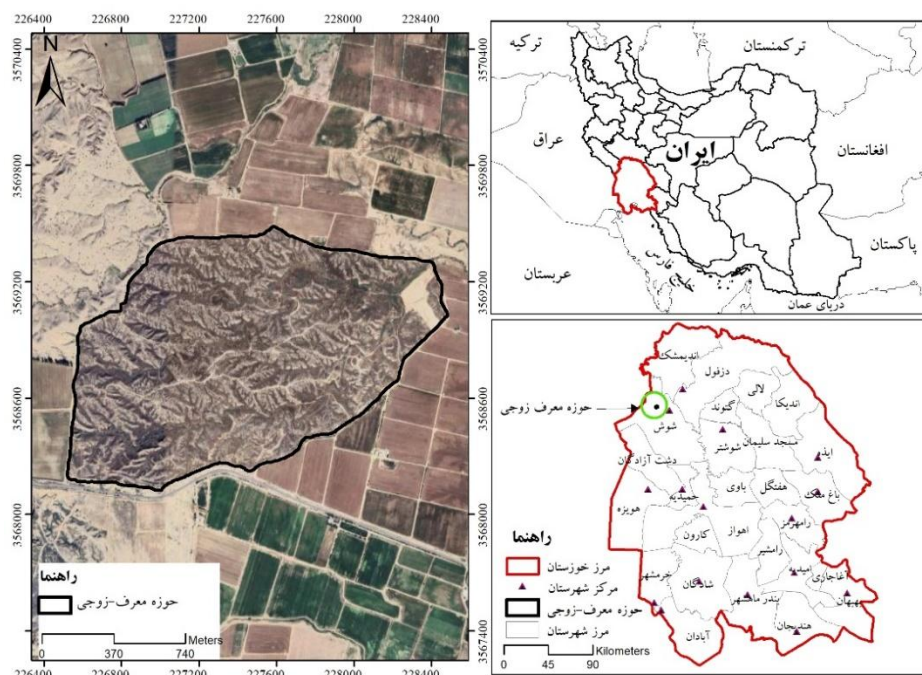
منطقه مورد مطالعه

این مطالعه در منطقه حفاظت‌شده آبخیز زوجی شوش، واقع در استان خوزستان انجام شده است (شکل ۱). این منطقه با مساحتی حدود ۲۰۰ هکتار و ارتفاعی بین ۹۰ تا ۱۴۷ متر از سطح دریا قرار دارد. نزدیک‌ترین منبع آب دائمی به منطقه، رودخانه کرخه است. اقلیم منطقه بر اساس روش آمبرژه به عنوان اقلیم گرم نیمه‌بیابانی طبقه‌بندی می‌شود. داده‌های هواشناسی ایستگاه سینوپتیک دزفول که نزدیک‌ترین ایستگاه به منطقه است، نشان می‌دهد میانگین بارش سالانه حدود ۳۰۰ میلی‌متر، با تمرکز در ماه‌های آذر تا اردیبهشت است.

1 Wadoux

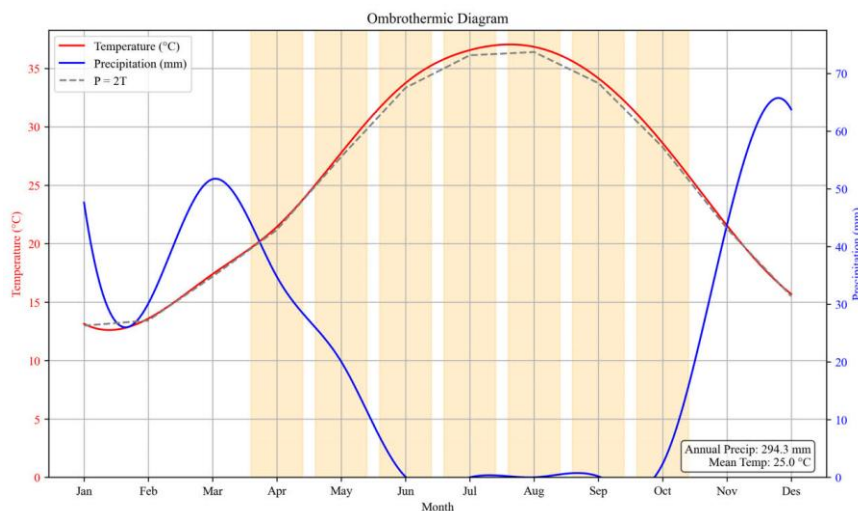
2 Dietze

3 Lundberg & Lee



شکل ۱. منطقه مورد مطالعه و نقاط نمونه برداری

دمای میانگین سالانه بین ۲۵ تا ۲۹ درجه سلسیوس متغیر است؛ ماههای تیر و مرداد گرمترین ماهها با حداکثر دمای مطلق ۴۵ تا ۵۰ درجه سلسیوس و میانگین ۳۴ تا ۳۶ درجه سلسیوس گرمترین دوره و دی ماه با دمای میانگین ۱۱ تا ۱۶ درجه سلسیوس نیز سردترین دوره است. بر اساس نمودار آمبروترمیک (شکل ۲)، منطقه در هفت ماه از سال با کمبود آب (بارش صفر یا حداقل) و دمای بالا مواجه است.



شکل ۲. نمودار آمبروترمیک منطقه مورد مطالعه

از نظر پوشش گیاهی، آبخیز مورد مطالعه در منطقه صحرا - سندی واقع شده و شامل تپه‌های کم‌ارتفاع و دره‌های باریک با خاک‌های با بافت لومی است. پوشش گیاهی عمدتاً شامل گونه‌های استپی کوتاه (گیاهان یک‌ساله و درختچه‌های خشک‌زی) است. این منطقه تنوع اکولوژیکی بالایی دارد که برای بررسی روابط غیرخطی گیاه در شرایط ناهمگون مناسب است.

روش تحقیق

طراحی نمونه برداری و جمع آوری داده ها

نمونه برداری میدانی در فصل بهار سال ۱۴۰۴ انجام شد. به منظور بررسی ارتباط ویژگی های خاک و پوشش گیاهی در واحدهای توپوگرافی مختلف، منطقه مورد مطالعه ابتدا به دو واحد اصلی شامل تپه^۱ و دره^۲ تفکیک شد. سپس در هر واحد توپوگرافی، نمونه برداری به روش ترانسکت بندی سیستماتیک انجام شد. برای این منظور، ترانسکت هایی به طول ۳۰ متر در سطح منطقه استقرار یافتند و فاصله بین ترانسکت های مجاور ۱۰ متر در نظر گرفته شد تا پوشش مناسبی از تغییرات مکانی منطقه فراهم شود. در امتداد هر ترانسکت، نقاط نمونه برداری به گونه ای انتخاب شدند که نماینده شرایط خاک و پوشش گیاهی همان واحد توپوگرافی باشند. در مجموع، ۱۲۰ نمونه شامل ۶۰ نمونه از واحد دره و ۶۰ نمونه از واحد تپه برداشت شد. موقعیت تمامی نقاط نمونه برداری در محیط GIS ثبت و کنترل شد تا توزیع مکانی نمونه ها در کل منطقه مورد مطالعه حفظ شود. این طرح نمونه برداری با هدف کاهش سوگیری مکانی، پوشش مناسب ناهمگونی های محیطی و فراهم کردن داده های قابل اعتماد برای مدل سازی روابط خاک - پوشش گیاهی و الگوریتم های یادگیری ماشین طراحی شد. متغیرهای اندازه گیری شده داده های خاک شامل ۹ ویژگی فیزیکی و شیمیایی ارائه شده در جدول ۱ و یک متغیر هدف (درصد پوشش گیاهی) بود.

جدول ۱. ویژگی های خاک اندازه گیری شده و روش های تجزیه آن ها

ویژگی خاک	نماد	روش اندازه گیری	واحد
اسیدیته خاک	pH	الکتروود شیشه ای در عصاره ۱:۲.۵ خاک-آب	-
فسفر قابل دسترس	P	روش Olsen	میلی گرم در کیلوگرم
پتاسیم تبادلی	K	استخراج با استات آمونیوم ۱ نرمال	میلی گرم در کیلوگرم
کربن آلی	OC	روش Walkley-Black	درصد
هدایت الکتریکی	EC	دستگاه هدایت سنج در عصاره اشباع	دسی زیمنس بر متر
ارزش خنثی سازی کل	TNV	تیتراسیون با اسید کلریدریک	درصد کربنات کلسیم معادل
درصد رس	Clay	روش هیدرومتری (پیپت)	درصد
درصد سیلت	Silt	روش هیدرومتری (پیپت)	درصد
درصد شن	Sand	روش هیدرومتری (پیپت)	درصد
پوشش گیاهی	VegCov	روش مستقیم	درصد

پیش پردازش داده ها

پس از جمع آوری، داده های خام مورد کنترل کیفیت قرار گرفتند. نمونه های فاقد مقدار برای متغیر هدف (پوشش گیاهی) حذف شدند. نرمال بودن توزیع داده ها با استفاده از آزمون Shapiro-Wilk بررسی شد. همچنین، برای مقابله با مقیاس پذیری متغیرها، تمامی ویژگی های ورودی با روش Standard Scaler نرمال سازی شدند.

تحلیل های آماری اولیه

برای توصیف داده ها از شاخص های آماری توصیفی شامل میانگین، میانه، انحراف معیار، کمینه و بیشینه استفاده شد. روابط بین متغیرها با ماتریس همبستگی ارزیابی شد. مقایسه های آماری با استفاده از آزمون های مناسب انجام گرفت (جدول ۲).

جدول ۲. آزمون‌های آماری مورد استفاده در مطالعه

نوع تحلیل	روش آماری	هدف	شرایط استفاده
بررسی نرمالیت	آزمون Shapiro-Wilk	بررسی توزیع نرمال داده‌ها	داده‌های پیوسته، حجم نمونه کوچک
همبستگی خطی	ضریب همبستگی پیرسون	شناسایی روابط خطی بین متغیرها	داده‌های نرمال، رابطه خطی
همبستگی غیرخطی	ضریب همبستگی اسپیرمن	شناسایی روابط یکنوا (خطی و غیرخطی)	داده‌های غیرنرمال یا ترتیبی
مقایسه دو گروه مستقل	آزمون t مستقل	مقایسه میانگین بین مناطق تپه و دره	داده‌های نرمال، واریانس برابر

مدل‌سازی یادگیری ماشین

برای پیش‌بینی درصد پوشش گیاهی بر اساس ۹ ویژگی‌های خاک، سه الگوریتم یادگیری ماشین مبتنی بر درخت انتخاب شدند. این انتخاب بر اساس توانایی آن‌ها در شناسایی روابط غیرخطی و آستانه‌ای (که در یافته‌های نتایج انعکاس یافته‌اند) انجام شد. پارامترهای بهینه هر مدل در جدول ۳ ارائه شده است.

جدول ۳. الگوریتم‌های یادگیری ماشین و پارامترهای بهینه آن‌ها

الگوریتم	پارامترهای کلیدی	مقادیر بهینه	توضیح
Random Forest (RF)	n_estimators, max_depth, min_samples_split	n_estimators=200, max_depth=10, min_samples_split=5	، مقاوم به ناهمگونی
XGBoost (XGB)	learning_rate, max_depth, n_estimators	learning_rate=0.1, max_depth=6, n_estimators=150	، روابط غیرخطی
Gradient Boosting (GB)	learning_rate, max_depth, n_estimators	learning_rate=0.1, max_depth=5, n_estimators=200	، تفسیرپذیر

داده‌ها به صورت طبقه‌بندی شده^۱ بر اساس واحد توپوگرافی به دو بخش آموزش (۷۰ درصد) و آزمون (۳۰ درصد) تقسیم شدند. به این ترتیب که از ۶۰ نمونه واحد تپه، ۴۲ نمونه (۷۰ درصد) به مجموعه آموزش و ۱۸ نمونه (۳۰ درصد) به مجموعه آزمون، و از ۶۰ نمونه واحد دره نیز به همین نسبت، ۴۲ نمونه به آموزش و ۱۸ نمونه به آزمون اختصاص یافت. این روش تضمین می‌کند که هر دو واحد توپوگرافی در هر دو مجموعه بازنمایی متوازن داشته باشند. برای جلوگیری از بیش‌برازش و ارزیابی پایایی مدل‌ها، از روش اعتبارسنجی متقابل پنج‌تایی طبقه‌بندی شده^۲ استفاده شد، به طوری که هر تا حاوی ۲۴ نمونه (۱۲ نمونه از تپه و ۱۲ نمونه از دره) بود. ر این روش داده‌های آموزش به ۵ قسمت مساوی تقسیم می‌شوند سپس مدل ۵ بار آموزش داده می‌شود (هر بار یک قسمت برای آزمون و ۴ قسمت برای آموزش) در آخر نتایج میانگین‌گیری می‌شوند. عملکرد مدل‌ها علاوه بر معیارهای R^2 و RMSE با شاخص‌های توزیعی شامل میانه، دامنه بین چارکی و انحراف معیار ارزیابی شد. با توجه به امکان ایجاد توزیع‌های نامتقارن و وجود مقادیر پرت در تکرارهای بوت‌استرپ، میانه R^2 به عنوان شاخص اصلی مقایسه عملکرد مدل‌ها مورد استفاده قرار گرفت. درخور یادآوری است که مقادیر منفی ضریب تعیین (R^2) نشان‌دهنده عملکرد ضعیف‌تر مدل نسبت به مدل پایه (میانگین داده‌ها) است. این پدیده در داده‌های با نویز بالا، روابط غیرخطی پیچیده و حجم نمونه محدود قابل وقوع است و فقط برای مقایسه نسبی مدل‌ها مورد استفاده قرار می‌گیرد (هاوکینز^۳، ۲۰۰۴).

تحلیل اهمیت ویژگی‌ها

اهمیت هر یک از ۹ متغیر خاک در پیش‌بینی پوشش گیاهی با استفاده از سه روش مکمل کمی‌سازی شد:

- 1 Stratified
- 2 Stratified 5-fold cross-validation
- 3 Hawkins

اهمیت ذاتی مدل بر پایه کاهش ناخالصی جینی: برای مدل های درختی بر اساس کاهش ناخالصی جینی در هر تقسیم محاسبه می شود (رابطه ۱):

$$I_{\text{Gini}}(f) = \sum_{\text{nodes}} P_{\text{node}} (1 - \sum_{j=1}^C p_{mj}^2) \quad (1)$$

F: ویژگی، Pnode: احتمال node، pmj: احتمال کلاس j در node، C: کلاس ها. اهمیت جایگشتی: این روش مدل مستقل است که با جابه جایی مقادیر هر ویژگی و مشاهده کاهش عملکرد مدل، اهمیت آن را به صورت مدل - مستقل محاسبه می کند (رابطه ۲).

$$PI(f) = \frac{Score_{\text{base}} - Score_{\text{perm}}(f)}{Score_{\text{base}}} \quad (2)$$

مقادیر SHAP: برای تفسیر سهم هر ویژگی در سطح نمونه های فردی و شناسایی روابط غیر خطی. درخور یادآوری است که مقادیر اهمیت ویژگی ها بسته به روش محاسبه می توانند متفاوت باشند. در این پژوهش، اهمیت ویژگی ها در شکل ۲-E مبتنی بر روش جایگشتی و در شکل ۳-A مبتنی بر کاهش ناخالصی جینی^۱ در مدل Gradient Boosting است. این تفاوت روش شناسی به رتبه بندی های متفاوت اما مکمل منجر می شود.

تحلیل عدم قطعیت و تفسیر پذیری مدل

برای ارزیابی قابلیت اطمینان پیش بینی ها و شناسایی منابع خطا، از روش های زیر استفاده شد: تخمین فاصله اطمینان: از روش بوت استرپ با ۵۰ تکرار برای تخمین فاصله اطمینان ۹۵ درصد پیش بینی ها و محاسبه انحراف معیار پیش بینی

شاخص های کیفی پیش بینی: برای این منظور از دو روش احتمال پوشش فاصله پیش بینی^۲: نسبت مشاهداتی که درون فاصله اطمینان ۹۵ درصد قرار می گیرند. عرض فاصله پیش بینی^۳: به صورت میانگین اختلاف بین کران بالا و پایین فاصله اطمینان ۹۵ درصد برای تمام نمونه های آزمون محاسبه شد. این شاخص بیانگر دقت و قطعیت پیش بینی ها است؛ هرچه عرض PI کوچک تر باشد، پیش بینی از قطعیت بیشتری برخوردار است.

در این مطالعه، عدم قطعیت کل پیش بینی ها با استفاده از روش بوت استرپ (با ۵۰ تکرار) به دو مؤلفه اصلی تفکیک شد: ۱- عدم قطعیت آلیاتوریک^۴: ناشی از تغییر پذیری ذاتی داده ها و خطاهای اندازه گیری است. این مؤلفه با استفاده از میانگین واریانس پیش بینی ها در تکرارهای بوت استرپ محاسبه می شود (رابطه ۳):

$$\sigma_{\text{aleatoric}}^2 = \frac{1}{N} \sum_{i=1}^N \mathbb{V}[y_i | x_i, \theta] \quad (3)$$

1 Gini Importance

2 PICP

3 PI Width

4 Aleatoric Uncertainty

۲- عدم قطعیت اپیستمیک^۱: ناشی از محدودیت دانش مدل (کمبود داده، انتخاب نادرست ساختار مدل، یا حذف متغیرهای مهم) است. این مؤلفه به عنوان واریانس پیش‌بینی‌ها بین تکرارهای مختلف بوت‌استرپ محاسبه می‌شود (رابطه ۴):

$$\sigma_{\text{epistemic}}^2 = \frac{1}{N} \sum_{i=1}^N (\mathbb{E}[y_i | x_i, \theta^{(k)}]) - \frac{1}{K} \sum_{k=1}^K \mathbb{E}[y_i | x_i, \theta^{(k)}])^2 \quad (4)$$

تحلیل زیرگروه‌ها و شرایط اکولوژیکی

برای بررسی اثر شرایط مختلف اکولوژیکی، داده‌ها به دو زیرمجموعه تقسیم و مدل‌سازی به صورت جداگانه انجام شد (جدول ۴). مدل‌سازی برای هر زیرمجموعه به صورت جداگانه انجام شد تا الگوهای خاص هر شرایط شناسایی شود. همچنین، برای شناسایی عوامل محدودکننده در مقیاس محلی، ۱۰ منطقه نمونه از بین نقاط نمونه‌برداری انتخاب و تحلیل منطقه‌ای روی آن‌ها انجام شد. برای شناسایی عوامل محدودکننده در مقیاس محلی، ۱۰ منطقه نمونه از بین ۱۲۰ نقطه نمونه‌برداری به صورت تصادفی - سیستماتیک انتخاب شدند. این مناطق معرف تنوع توپوگرافی و پوشش گیاهی منطقه مورد مطالعه بودند. در هر منطقه، میانگین ویژگی‌های خاک و پوشش گیاهی محاسبه و فراوانی عوامل محدودکننده تعیین شد.

جدول ۴. زیرمجموعه‌های مورد تحلیل در مطالعه

زیر مجموعه	توضیح	تعداد نمونه	هدف تحلیل
مناطق تپه‌ای	تمام نمونه‌های جمع‌آوری شده از واحد توپوگرافی تپه	60	شناسایی الگوهای خاص مناطق مرتفع
مناطق دره‌ای	تمام نمونه‌های جمع‌آوری شده از واحد توپوگرافی دره	60	شناسایی الگوهای خاص مناطق پست

پیاده‌سازی نرم‌افزاری

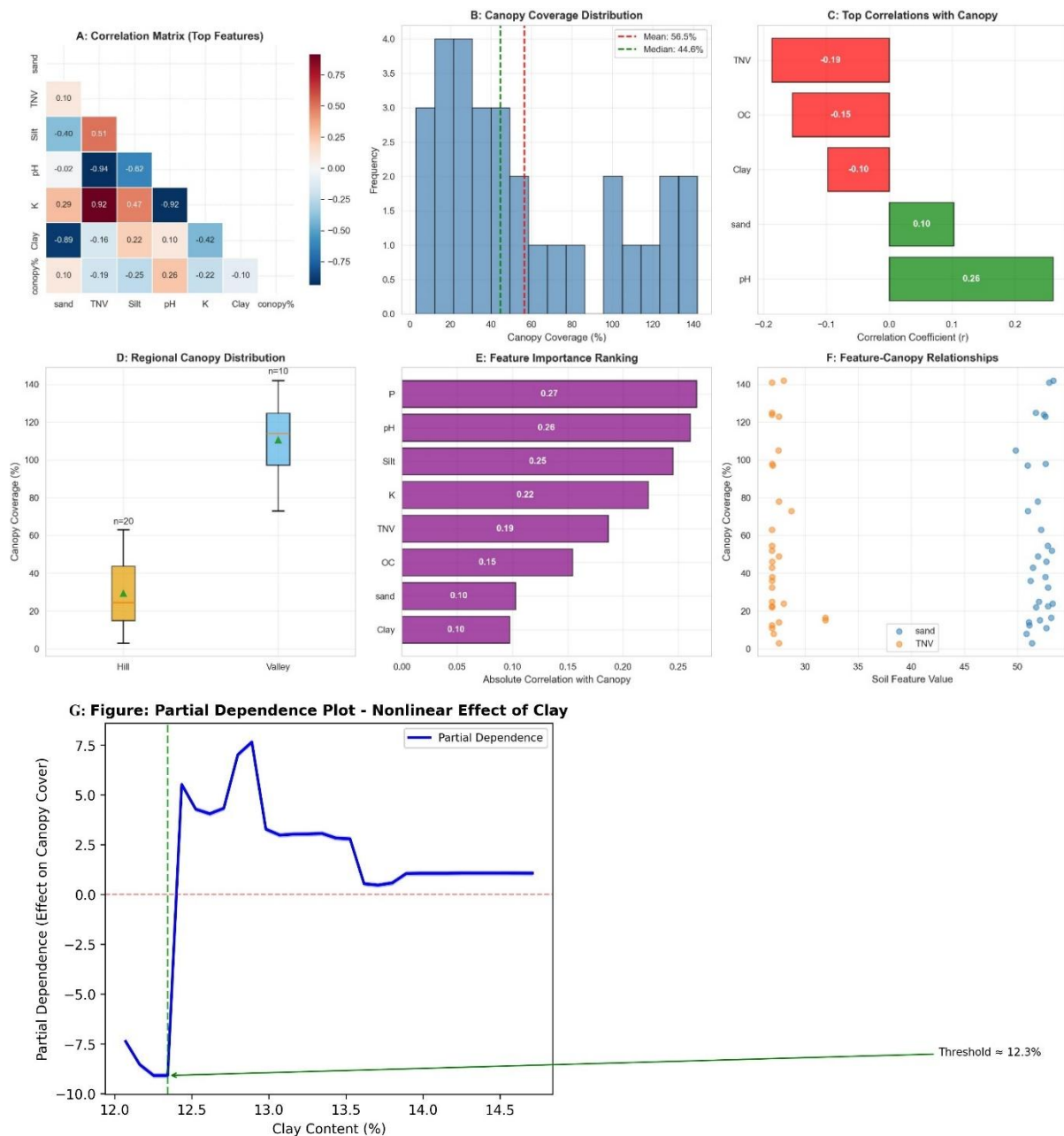
تمامی مراحل تحلیل داده، مدل‌سازی و مصورسازی با استفاده از زبان برنامه‌نویسی پایتون (نسخه ۳.۱۱) و کتابخانه‌های تخصصی زیر انجام شد:

- Scikit-learn (نسخه ۱.۳۰) برای الگوریتم‌های پایه یادگیری ماشین
- XGBoost (نسخه ۱.۷.۶) برای پیاده‌سازی الگوریتم XGBoost
- SHAP (نسخه ۰.۴۲.۱) برای تحلیل تفسیرپذیری مدل
- Pandas (نسخه ۲.۰.۳) و NumPy (نسخه ۱.۲۴.۳) برای پردازش داده
- Matplotlib (نسخه ۳.۷.۱) و Seaborn (نسخه ۰.۱۲.۲) برای ترسیم نمودارها.

یافته‌های پژوهش

تحلیل روابط خاک-گیاه

توزیع پوشش تاج‌پوشش (شکل ۳-B) با میانگین ۵۶/۵ درصد و میانه ۴۴/۶ درصد، یک چولگی مثبت واضح را نشان می‌دهد. نتایج ماتریس همبستگی پیرسون (جدول ۵ و شکل ۳-A) نشان داد روابط خطی بین ویژگی‌های خاک و درصد تاج‌پوشش عموماً ضعیف تا متوسط هستند. بیشترین همبستگی مثبت با تاج‌پوشش مربوط به pH (۰/۲۶۱/ r) و بیشترین همبستگی منفی مربوط به فسفر قابل جذب (۰/۲۶۷/ r) بود. با این حال، هیچ یک از متغیرهای خاکی همبستگی خطی قوی با پوشش گیاهی نشان ندادند. این یافته بیانگر آن است که تغییرات پوشش گیاهی احتمالاً تحت تأثیر برهم‌کنش همزمان چندین عامل خاکی و روابط غیرخطی بین متغیرها قرار دارد. بنابراین، نتایج جدول ۵ ضرورت استفاده از رویکردهای چندمتغیره و الگوریتم‌های یادگیری ماشین را برای شناسایی الگوهای پنهان در روابط خاک - پوشش گیاهی تأیید می‌کند.



بالا ترین C: توزیع پوشش تاج پوشش؛ B: ماتریس همبستگی (ویژگی‌های برتر خاک)؛ شکل ۳. A: الگوهای کلیدی آماری در روابط خاک - پوشش گیاهی؛
روابط کمی بین ویژگی‌های F: رتبه‌بندی اهمیت ویژگی‌ها در مدل پیش‌بینی؛ E: توزیع منطقه‌ای پوشش و عوامل محدودکننده D همبستگی‌ها با پوشش تاج پوشش؛
اثر غیرخطی مقدار رس بر پوشش تاجی و شناسایی نقطه آستانه در حدود ۱۲/۳ درصد G خاک و پوشش تاج پوشش؛

توزیع پوشش تاج پوشش (شکل B-۳) با میانگین ۵۶/۵ درصد و میان ۴۴/۶ درصد، یک چولگی مثبت واضح را نشان می‌دهد. این اختلاف قابل توجه به این معناست که اگرچه بیشتر منطقه پوشش گیاهی نسبتاً کم‌تراکم دارد، اما وجود نقاط خاص (در کرانه رودخانه‌ها) با پوشش بسیار متراکم، میانگین کلی را بالا برده است. این الگوی ناهمگن، لزوم در نظرگیری عوامل میکروتوپوگرافی و مدیریتی را در کنار ویژگی‌های خاکی تأکید می‌کند، موضوعی که در مطالعات مقیاس‌گذاری اکوسیستم‌های ناهمگون همواره مورد توجه است. در بررسی همبستگی مستقیم خاک با پوشش گیاهی (شکل C-۳)، بر اساس ضریب همبستگی پیرسون، هیچ‌یک از ویژگی‌های خاک همبستگی خطی معناداری با پوشش تاج پوشش نشان ندادند ($p\text{-value} > 0.05$). با این حال، بر اساس ضریب

همبستگی اسپیرمن، سیلت با همبستگی منفی -0.4 ($p < 0.05$) قوی‌ترین رابطه یکنوا (و منفی) را با پوشش گیاهی داشت. TNV (-0.19) و مواد آلی (-0.15) نیز همبستگی منفی (اما غیرمعنادار آماری) نشان دادند. در مقابل، شن ($+0.10$) و pH ($+0.26$) همبستگی مثبت ضعیفی داشتند. این نتایج نشان می‌دهد شیوه‌های آماری خطی قادر به کشف روابط پنهان بین خاک و پوشش گیاهی نیستند و ضرورت استفاده از مدل‌های یادگیری ماشین را تأیید می‌کند.

مطالعه کائو و همکاران (۲۰۱۹) در مناطق نیمه‌خشک نیز نشان داد ویژگی‌های فیزیکی خاک (به‌ویژه بافت خاک و ظرفیت نگهداری آب) نقش تعیین‌کننده‌ای در تغییرات پوشش گیاهی دارند، اما این روابط اغلب غیرخطی و آستانه‌ای هستند.

لازم به توضیح است که رتبه‌بندی متغیرها بسته به روش تحلیل متفاوت است. تحلیل منطقه‌ای (شکل ۳-D) عوامل محدودکننده را در مقیاس محلی و بر اساس فراوانی نسبی در ۱۰ منطقه شناسایی می‌کند، در حالی که روش Gini Importance (شکل ۴-A) کاهش ناخالصی حاصل از هر متغیر را در مدل Gradient Boosting اندازه‌گیری می‌کند. در مقابل، تحلیل SHAP (شکل ۶L-J) یک معیار مدل آگنوستیک است که سهم هر متغیر را در پیش‌بینی‌های مدل محاسبه می‌کند. با توجه به مزایای روش SHAP (عدم وابستگی به الگوریتم خاص، مقاومت در برابر هم‌خطی متغیرها، و قابلیت تفسیر محلی)، این روش به عنوان مبنای نتیجه‌گیری نهایی در این مطالعه انتخاب شده است. بر این اساس، مواد آلی، درصد شن و TNV به عنوان مهم‌ترین عوامل کنترل‌کننده پوشش گیاهی در منطقه مورد مطالعه شناسایی می‌شوند. قابل توجه است که اگرچه فسفر در تحلیل منطقه‌ای به عنوان مهم‌ترین عامل محدودکننده در ۷۵ درصد مناطق شناسایی شد، در تحلیل SHAP در رتبه‌های پایین‌تری قرار گرفت. این تفاوت نشان‌دهنده آن است که تأثیر فسفر بر پوشش گیاهی ممکن است بیشتر در مقیاس محلی (منطقه‌ای) قابل شناسایی باشد تا در سطح مدل کلی. با این حال، در خاک‌های آهکی، معمولاً تصور می‌شود اصلاح pH اولویت دارد، اما داده‌های منطقه‌ای این مطالعه نشان می‌دهند محدودیت فسفر حتی در بازه‌های pH نسبتاً قابل تحمل نیز می‌تواند عامل مسلط باشد. این یافته با کار هینزینگر (۲۰۰۱) همخوانی دارد که بر «تله فسفاتی» در خاک‌های غنی از کلسیم تأکید کرد. بنابراین، راهبردهای احیا باید مستقیم بر بهبود چرخه و در دسترس بودن فسفر (مثلاً از طریق کوددهی هدفمند، استفاده از میکوریزا یا کشت گیاهان آزادکننده فسفر) متمرکز شوند، نه فقط بر کاهش pH.

جالب توجه است که مواد آلی خاک در ماتریس همبستگی ساده (شکل ۳-C) رابطه منفی ضعیفی با پوشش گیاهی نشان داد، اما در رتبه‌بندی اهمیت ویژگی‌ها در مدل پیش‌بینی (شکل ۶L-J) در جایگاه قابل توجهی قرار گرفت. این احتمال وجود دارد که در حضور متغیرهای محدودکننده‌ای مانند رس یا آهک، اثر مستقیم مثبت خود را آشکار نکند، اما در یک مدل چندمتغیره که اثرات این متغیرها کنترل می‌شوند، نقش مفید آن در بهبود ساختمان خاک، نگهداری آب و تغذیه گیاه مشخص شود. همچنین، ممکن است رابطه مواد آلی خاک با پوشش گیاهی غیرخطی باشد، پدیده‌ای که در همبستگی خطی ساده قابل مشاهده نیست اما در مدل‌های پیشرفته قابل شناسایی است.

نمودار روابط کمی (شکل ۳-F) به طور عینی روند غیرخطی تأثیر متغیرهای خاک بر پوشش گیاهی را نشان می‌دهد. این نمودار رابطه افزایشی ملایم پوشش گیاهی با افزایش برخی متغیرها تا یک آستانه و رابطه کاهشی با افزایش برخی دیگر را مستقیم از داده‌ها استخراج کرده و بحث‌های کیفی فوق را پشتیبانی می‌کند. وجود این آستانه‌ها نکته‌ای حیاتی برای مدیریت است: بهبود بافت خاک (مثلاً با افزودن ماده آلی یا شن) تنها تا یک نقطه مشخص سودمند خواهد بود و برنامه‌ریزی‌ها باید بر اساس این نقاط بهینه انجام شود.

جدول ۵. ضرایب همبستگی پیرسون بین درصد تاج پوشش و ویژگی های خاک

متغیر	ضریب همبستگی (r)
pH	۰/۲۶۱
P	-۰/۲۶۷
Silt	-۰/۲۴۵
K	-۰/۲۲۳
EC	-۰/۲۱۴
TNV	-۰/۱۸۷
OC	-۰/۱۵۴
Sand	۰/۱۰۳
Clay	-۰/۰۹۸

افزون بر تحلیل همبستگی و رتبه بندی متغیرها، بررسی اثرات غیرخطی رس با استفاده از نمودار وابستگی جزئی (شکل ۳-G) بینش تکمیلی مهمی فراهم می کند. این نمودار نشان می دهد اثر رس بر تاج پوشش ماهیتی آستانه ای دارد؛ به گونه ای که افزایش رس تا حدود ۱۲/۳ درصد با کاهش تند پوشش گیاهی همراه است، اما پس از این نقطه، شدت اثر به طور قابل ملاحظه ای کاهش می یابد. این رفتار با یافته های گزارش شده برای اکوسیستم های نیمه خشک همخوان است که در آن ها تغییرات جزئی در بافت خاک می تواند منجر به جهش های ناگهانی در محدودیت های آبی و فیزیکی شود.

این نتیجه همچنین تبیین می کند که چرا رس، با وجود همبستگی خطی ضعیف با پوشش گیاهی، در مدل های یادگیری ماشین نقش غیرخطی معناداری ایفا می کند. به طور مدیریتی، حضور این آستانه نشان می دهد اصلاحات فیزیکی خاک (نظیر بهبود ساختمان یا افزایش ماده آلی) عمدتاً در خاک هایی با رس کمتر از ۱۲/۳ درصد مؤثرند، در حالی که در خاک های رسی تر، اولویت با مداخله های تغذیه ای و رطوبتی است.

تحلیل جامع اهمیت ویژگی های خاک در پیش بینی پوشش گیاهی

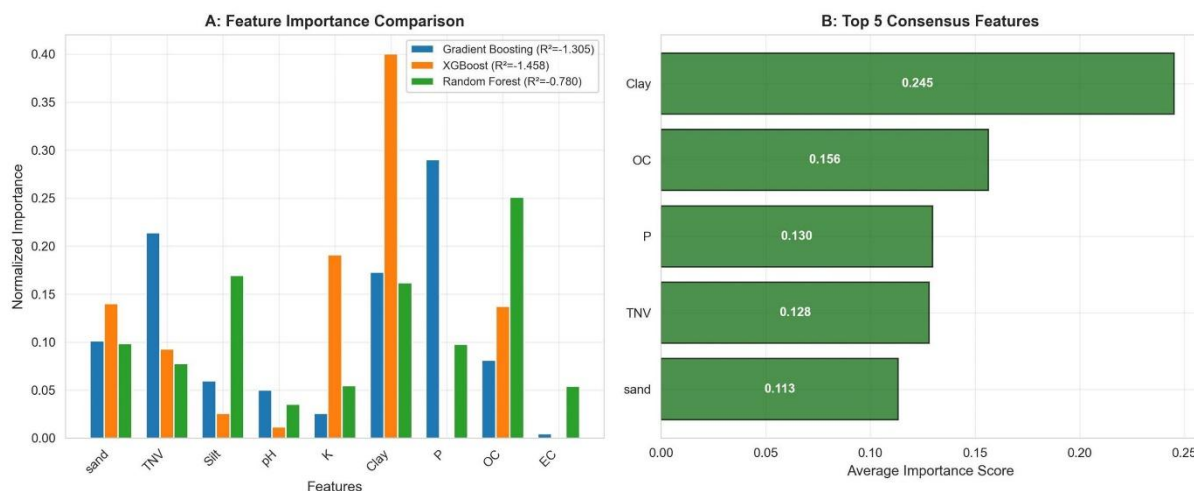
برای شناسایی دقیق ترین عوامل کنترل کننده تاج پوشش و عبور از محدودیت های تحلیل های همبستگی ساده، از روش های یادگیری ماشین مبتنی بر درخت^۱ استفاده شد. این رویکرد قادر است روابط غیرخطی، تعامل های پیچیده بین متغیرها و اهمیت واقعی هر پیش بینی در یک چارچوب چندمتغیره را آشکار کند.

با توجه به اینکه مدل سازی اولیه بر روی داده های یکپارچه (بدون تفکیک توپوگرافی) به دلیل ناهمگونی محیطی و حجم نمونه محدود (۳۰ نمونه) منجر به عملکرد ضعیف مدل ها شد، در ادامه تحلیل روی زیرمجموعه های همگن (تپه و دره) متمرکز شد. نتایج مدل سازی تفکیکی در شکل ۵ ارائه شده است که عملکرد قابل قبول مدل ها را در منطقه دره ($R^2 \approx 0.75$) نشان می دهد. این یافته تأکید می کند که نادیده گرفتن ناهمگونی توپوگرافی و حجم نمونه ناکافی می تواند به شکست مدل های پیچیده یادگیری ماشین منجر شود. با این حال، مقایسه داخلی مدل ها، برتری نسبی Gradient Boosting را در پردازش این مجموعه داده خاص تأیید می کند. این الگوریتم با تکنیک تقویت تدریجی^۲ که روی خطاهای باقی مانده مدل های ضعیف قبلی تمرکز می کند، ظاهراً توانسته است الگوهای پیچیده تری را در مقایسه با رویکرد bagging در Random Forest بیاموزد. این یافته با نتایج پژوهش حیدری و همکاران (۱۴۰۴) در زمینه کاربرد الگوریتم های یادگیری ماشین در پیش بینی ویژگی های خاک همخوانی دارد. آن ها نیز نشان دادند مدل های مبتنی بر درخت، به ویژه در داده های با حجم محدود و روابط غیرخطی، عملکرد بهتری نسبت به مدل های پیچیده دارند. نتایج حاضر از منظر نوع روابط بین متغیرها نیز با یافته های محمدیان و همکاران (۲۰۲۵) و حیدری و همکاران (۲۰۲۶) همخوانی دارد. در آن پژوهش ها نشان داده شد که متغیرهای خاکی اغلب از طریق برهم کنش های پیچیده و غیرخطی بر

1 Tree-based Ensemble Models

2 Boosting

پاسخ‌های محیطی اثر می‌گذارند و به همین دلیل الگوریتم‌های مبتنی بر درخت تصمیم قادرند بخش بیشتری از تغییرات را نسبت به روش‌های آماری کلاسیک تبیین کنند. در پژوهش حاضر نیز نبود همبستگی‌های خطی قوی در کنار عملکرد مناسب مدل‌های یادگیری ماشین، بیانگر آن است که روابط خاک - پوشش گیاهی عمدتاً ماهیتی غیرخطی و وابسته به اثرات متقابل متغیرها دارند.



شکل ۴. تحلیل جامع اهمیت ویژگی‌های خاک در پیش‌بینی پوشش گیاهی. A: مقایسه عملکرد مدل‌ها و اهمیت نرمال‌شده ویژگی‌ها؛ B: پنج ویژگی برتر اجماعی (توافق‌شده بین مدل‌ها)

تحلیل اهمیت نرمال‌شده ویژگی‌ها (شکل A-۴) که بر پایه کاهش ناخالصی^۱ در مدل Gradient Boosting محاسبه شده، یک رتبه‌بندی متفاوت و عمیق‌تر از عوامل مؤثر را نسبت به تحلیل همبستگی ساده بخش قبلی (شکل E-۳) ارائه می‌دهد. پنج ویژگی برتر به ترتیب عبارتند از: رس با اهمیت ۰/۲۴۵، ماده آلی خاک با ۰/۱۵۶، فسفر با ۰/۱۳، ارزش خنثی‌سازی (TNV) با ۰/۱۲۸ و شن با ۰/۱۱۳. شکل B-۴ این پنج عامل برتر را به عنوان عوامل اجماع^۲ برجسته می‌سازد. این رتبه‌بندی چندین نکته کلیدی را آشکار می‌سازد که مستلزم بحثی عمیق‌تر است.

نخست، نقش مسلط رس به عنوان مهم‌ترین پیش‌بین در مدل، در تضاد ظاهری با همبستگی خطی ضعیف و منفی آن با پوشش گیاهی ($r = -0.10$ در شکل E-۳) قرار دارد. این تناقض نشان‌دهنده محدودیت تحلیل همبستگی خطی و توانایی مدل‌های درختی در آشکارسازی روابط غیرخطی است. به منظور بررسی دقیق این رابطه، نمودار وابستگی جزئی رس (PDP؛ شکل G-۳) تهیه شد. نتایج نشان داد که اثر رس بر تاج پوشش دارای رفتار آستانه‌ای است؛ به طوری که افزایش رس تا ۱۲/۳ درصد با کاهش نسبتاً شدید پوشش گیاهی همراه بوده، اما پس از این آستانه به شدت اثر کاهش می‌یابد. این الگو احتمالاً ناشی از تغییرات ناگهانی در ویژگی‌های فیزیکی خاک، از جمله کاهش تهویه، محدودیت نفوذ ریشه و تغییر رژیم رطوبتی است. بنابراین، مدل Gradient Boosting توانسته است رابطه غیرخطی و آستانه‌ای رس را شناسایی کند؛ رابطه‌ای که با روش‌های خطی کلاسیک قابل آشکارسازی نبود. این یافته با مطالعات پاداریان و همکاران (۲۰۲۰) همخوانی دارد که نشان دادند مدل‌های یادگیری ماشین قادر به شناسایی اثرات غیرخطی و وابسته به دامنه متغیرهای خاک هستند. دوم، اهمیت ماده آلی خاک نه تنها به عنوان یک منبع غذایی، بلکه به دلیل بهبود ساختار خاک، نفوذپذیری آب، ظرفیت تبادل کاتیونی و فعالیت میکروبی، می‌تواند اثرات منفی محتوای بالای رس و آهک (TNV) را تعدیل کند. سوم، تأیید مجدد فسفر به عنوان عوامل کلیدی شیمیایی محدودکننده است که با نتایج تحلیل منطقه‌ای شکل E-۳ همخوانی دارد. حضور قوی این متغیر در بین پنج عامل برتر مدل پیشرفته، یافته‌های اولیه را اعتبار

1 Gini Importance

2 Consensus Features

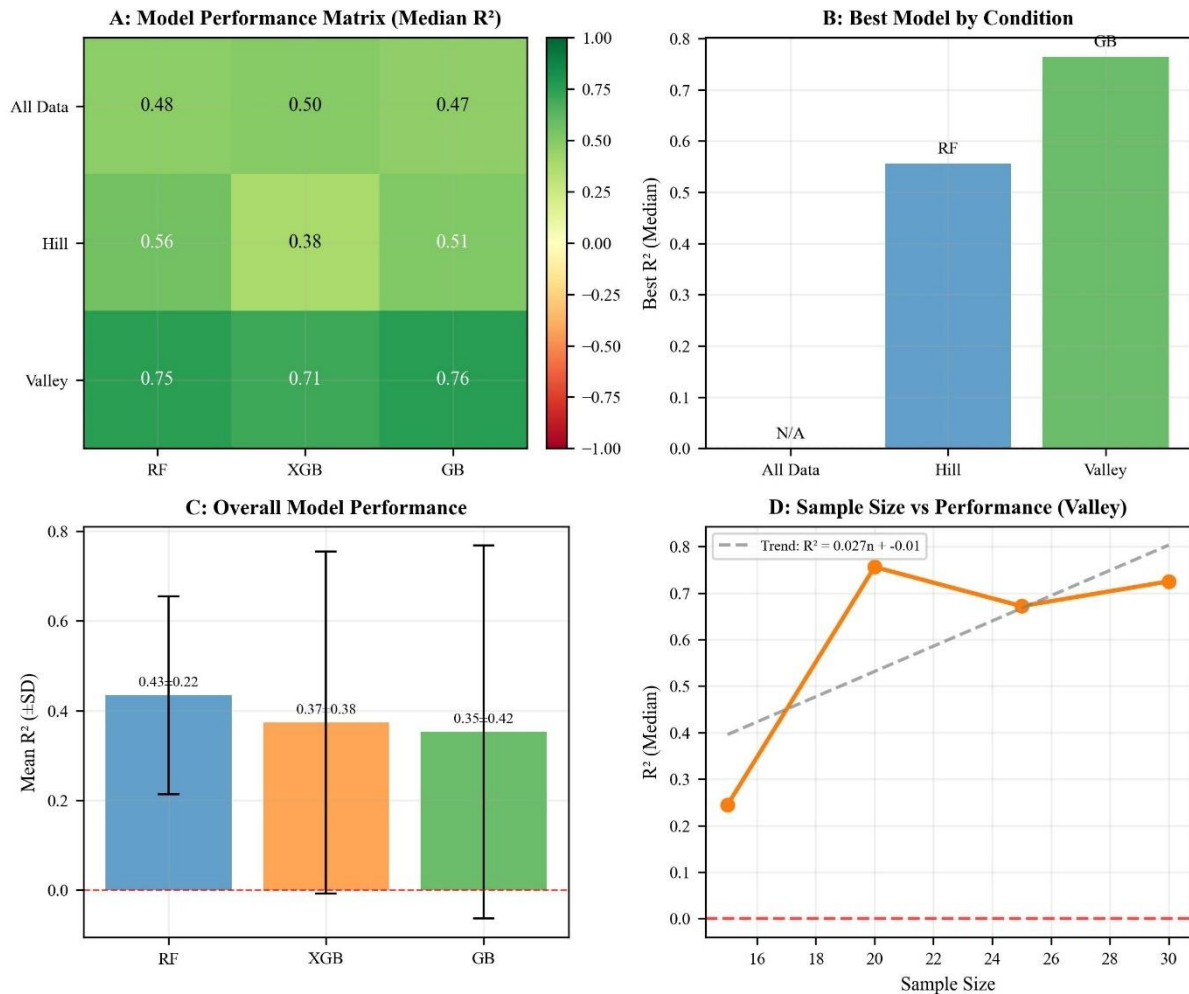
می‌بخشد و بر این واقعیت تأکید می‌کند که محدودیت فسفر در خاک‌های آهکی یک عامل تعیین‌کننده در مقیاس مدل‌سازی منطقه‌ای است. این نتیجه با مطالعات اخیر انجام‌شده در زمینه کاربرد یادگیری ماشین در علوم خاک نیز همخوان است. حیدری و همکاران (۲۰۲۶) گزارش کردند که متغیرهای مرتبط با حاصلخیزی خاک، به‌ویژه عناصر غذایی و ماده آلی، سهم قابل توجهی در قدرت پیش‌بینی مدل‌های داده‌محور دارند. در پژوهش حاضر نیز حضور رس، ماده آلی و فسفر در میان مهم‌ترین متغیرهای پیش‌بینی‌کننده، اهمیت کیفیت و وضعیت تغذیه‌ای خاک را به عنوان عامل کلیدی کنترل‌کننده پوشش گیاهی تأیید می‌کند. مقایسه نتایج جدول ۵ با تحلیل اهمیت متغیرها نشان می‌دهد شدت همبستگی خطی یک متغیر الزاماً بیانگر اهمیت آن در مدل‌های یادگیری ماشین نیست. به عنوان مثال، فسفر ($r = -0.267$)، ماده آلی ($r = -0.154$) و رس ($r = -0.098$) اگرچه همبستگی خطی نسبتاً ضعیفی با درصد تاج‌پوشش نشان دادند (جدول ۵)، اما در تحلیل اهمیت ویژگی‌ها و مقادیر SHAP در زمره متغیرهای کلیدی قرار گرفتند. این تفاوت ناشی از ماهیت الگوریتم‌های یادگیری ماشین است که علاوه بر روابط خطی، قادر به شناسایی اثرات غیرخطی، تعامل بین متغیرها و پاسخ‌های آستانه‌ای هستند. بنابراین، نتایج حاصل از جدول ۵ و تحلیل‌های مبتنی بر یادگیری ماشین نه تنها متناقض نیستند، بلکه مکمل یکدیگر در تبیین روابط پیچیده خاک - پوشش گیاهی محسوب می‌شوند.

ارزیابی جامع و مقایسه عملکرد مدل‌های پیش‌بینی پوشش گیاهی در چارچوب محدودیت‌های خاک

پس از شناسایی ویژگی‌های کلیدی خاک (شکل ۴) و الگوهای اولیه ارتباط آنها با پوشش گیاهی (شکل ۳)، مرحله بعدی ارزیابی سیستماتیک مدل‌های پیش‌بینی است تا مشخص شود کدام الگوریتم قادر است این روابط پیچیده و اغلب غیرخطی را با بیشترین دقت و پایدارترین حالت بازتولید کند. شکل ۵ نتایج این تحلیل چندبعدی را ارائه می‌دهد و بینش‌های عمیقی در مورد قابلیت اطمینان مدل‌سازی در محیط ناهمگون منطقه مورد مطالعه فراهم می‌آورد.

مقایسه کلی عملکرد سه الگوریتم پیشرو (شکل ۵-A) نشان می‌دهد داده‌های ترکیبی (All Data)، مدل XGBoost با میانۀ R^2 ۰/۵۸ عملکرد برتر نسبت به Gradient Boosting با میانۀ عملکرد ۰/۴۷، و Random Forest با میانۀ عملکرد ۰/۳۷ دارد. با این حال، در تفکیک بر اساس واحد توپوگرافی، عملکرد مدل‌ها در منطقه دره (Valley) به طور قابل توجهی بالاتر است، به طوری که XGBoost و Gradient Boosting هر دو به میانۀ R^2 حدود ۰/۷۱ و ۰/۷۶ دست می‌یابند. این یافته نشان می‌دهد مدل‌های مبتنی بر بوستینگ به‌ویژه در زیرمجموعه‌های همگن محیطی توانایی بالایی در شبیه‌سازی روابط خاک - گیاه دارند.

Figure 4: Comprehensive Model Evaluation



شکل ۵. ارزیابی جامع و مقایسه عملکرد مدل‌های پیش‌بینی پوشش گیاهی در شرایط مختلف. A: مقایسه توزیع عملکرد سه مدل پیشرو (Gradient Boosting, Random Forest, XGBoost); B: شناسایی مدل بهینه تحت شرایط مختلف (منطقه دره، تپه، و کل داده‌ها); C: عملکرد کلی مدل‌های برتر با شاخص‌های آماری (میانگین $\pm$ انحراف معیار); D: کاهش بازده افزایشی «Diminishing Returns».

انتخاب بهترین مدل تحت شرایط مختلف (شکل B-۵) تحلیل را به سطحی کاربردی‌تر می‌برد. نتایج نشان می‌دهد در منطقه دره، هر سه مدل عملکرد قابل قبولی دارند (XGBoost با R^2 نزدیک به یک بهترین عملکرد را نشان می‌دهد)، در حالی که در منطقه تپه، Random Forest با R^2 حدود ۵۰ عملکرد نسبی بهتری دارد. این تنوع در عملکرد مدل‌ها در واحدهای توپوگرافی مختلف، نشان‌دهنده وابستگی شدید روابط خاک - گیاه به بافت محیطی است و لزوم مدل‌سازی تفکیکی را تأکید می‌کند. عملکرد کلی مدل‌های برتر با شاخص‌های آماری (شکل C-۵) نشان می‌دهد میانگین R^2 در داده‌های ترکیبی برای XGBoost حدود ۰/۳۷، برای Random Forest حدود ۰/۴۴ و برای Gradient Boosting حدود ۰/۳۵ است. هرچند Forest Random در برخی زیرنمونه‌ها پایداری بیشتری نشان داد، اما بر اساس میانه R^2 ، مدل XGBoost عملکرد کلی بهتری ارائه کرد. این یافته بیانگر آن است که در مجموعه داده ترکیبی (شامل هر دو واحد توپوگرافی)، Random Forest به دلیل مکانیسم بگینگ^۱ و مقاومت نسبی در برابر ناهمگونی داده‌ها، عملکردی پایدارتر و قابل اعتمادتر از مدل‌های بوستینگ دارد.

این یافته برای پژوهش ما حیاتی است و پیشنهاد می‌کند که ایجاد مدل‌های جداگانه برای واحدهای همگن اکولوژیکی - پدوژنتیکی (مانند واحدهای تپه‌ای و دره‌ای) می‌تواند دقت پیش‌بینی را افزایش دهد. مطالعات قبلی نیز نشان داده‌اند که در داده‌های با روابط غیرخطی قوی، مدل‌های بوستینگ اغلب از مدل‌های بگینگ پیشی می‌گیرند (شوارتز و همکاران، ۲۰۱۹). تأثیر حجم نمونه بر عملکرد مدل شکل (شکل ۵-D) یکی از یافته‌های کلیدی و کاربردی این پژوهش است. عملکرد مدل در منطقه دره با افزایش حجم نمونه از ۲۰ به ۶۰ نمونه، از حدود ۰/۷ به ۰/۸ بهبود یافته و سپس در حجم‌های بالاتر روند افزایشی آن با شیب کمتری ادامه می‌یابد. البته تأیید قطعی این آستانه نیازمند مطالعات آتی با حجم نمونه بیشتر و طراحی subsampling منظم است (وادکس و همکاران، ۲۰۱۹).

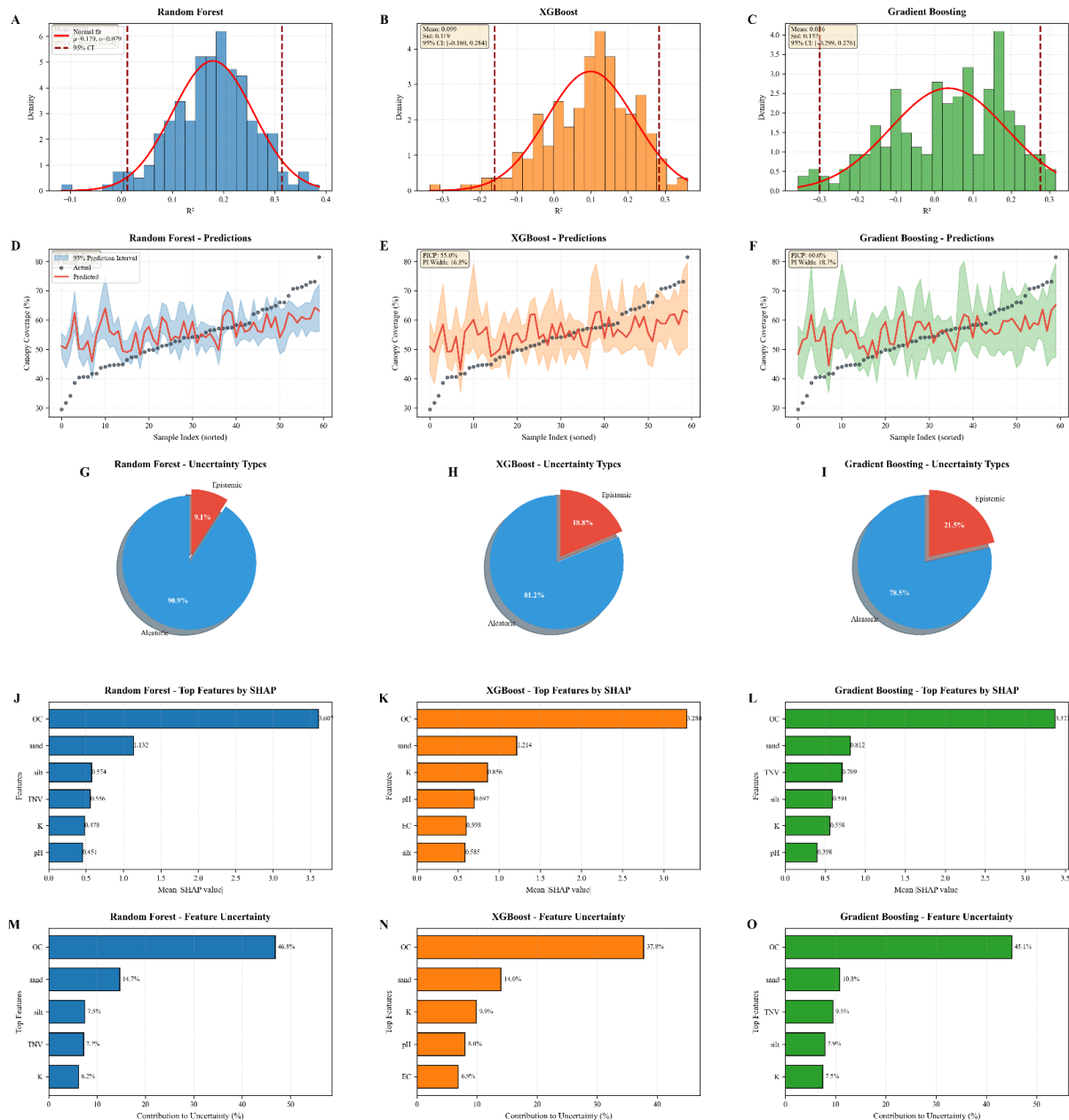
به طور کلی، یافته‌های این بخش برای پژوهش ما حیاتی است و پیشنهاد می‌کند که ایجاد مدل‌های جداگانه برای واحدهای همگن اکولوژیکی - پدوژنتیکی (مانند واحدهای تپه‌ای و دره‌ای) می‌تواند دقت پیش‌بینی را به طور معناداری افزایش دهد. مطالعات قبلی نیز نشان داده‌اند در داده‌های با روابط غیرخطی قوی و ناهمگونی محیطی، انتخاب الگوریتم مناسب باید بر اساس ویژگی‌های زیرمجموعه داده‌ها انجام شود (شوارتز و همکاران، ۲۰۱۹). یافته‌های حاضر همچنین با نتایج پژوهش‌های محمدیان و همکاران (۲۰۲۵) در زمینه مدل‌سازی ویژگی‌های خاک و محیط با استفاده از الگوریتم‌های یادگیری ماشین مطابقت دارد. در آن مطالعات نیز نشان داده شد که تفکیک داده‌ها به زیرگروه‌های همگن و توجه به ناهمگنی محیطی می‌تواند بهبود قابل توجهی در عملکرد مدل‌ها ایجاد کند. افزایش چشمگیر دقت مدل‌ها در واحد دره نسبت به داده‌های ترکیبی در پژوهش حاضر نیز مؤید اهمیت لحاظ کردن ساختار اکولوژیکی و توپوگرافی در فرآیند مدل‌سازی است.

عدم قطعیت و حساسیت مدل‌های پیش‌بینی

تحلیل عملکرد و ارزیابی عدم قطعیت مدل‌های یادگیری ماشین گامی حیاتی در تضمین قابلیت اطمینان و کاربردپذیری پیش‌بینی‌های اکولوژیکی است. در این بخش، سه مدل برجسته Random Forest (RF)، XGBoost (XGB) و Gradient Boosting (GB) با استفاده از رویکردهای ترکیبی بوت‌استرپ، تحلیل SHAP و کمی‌سازی عدم قطعیت، به‌طور جامع مورد ارزیابی قرار گرفته‌اند. نتایج این تحلیل چندبعدی در شکل ۶ ارائه شده که شامل ۱۵ زیربخش در ۵ ردیف موضوعی است.

ارزیابی عملکرد و پایداری مدل‌ها

تحلیل توزیع مقادیر R^2 حاصل از تکرارهای بوت‌استرپ (شکل ۶A-C)، درجه پایداری و حساسیت مدل‌ها را نشان می‌دهد. با توجه به توزیع نامتقارن مقادیر R^2 ، در این مطالعه میانه به عنوان شاخص اصلی عملکرد مدل‌ها در نظر گرفته شد. نتایج نشان داد XGBoost دارای بالاترین میانه عملکرد در داده‌های ترکیبی ($R^2=0/5$) بوده است، در حالی که Gradient Boosting و Random Forest به ترتیب مقادیر ۰/۴۷ و ۰/۳۷ را نشان دادند. با این حال، پراکندگی نتایج بوت‌استرپ بیانگر آن بود که عملکرد مدل‌ها در زیرنمونه‌های مختلف با تغییراتی همراه است که ناشی از ناهمگنی محیطی و پیچیدگی روابط خاک - پوشش گیاهی است. وجود مقادیر منفی R^2 در برخی تکرارهای اعتبارسنجی متقابل و بوت‌استرپ نشان می‌دهد در آن تکرارها عملکرد مدل از مدل مبنا (پیش‌بینی میانگین متغیر هدف) ضعیف‌تر بوده است. چنین وضعیتی در مجموعه داده‌های دارای حجم نمونه محدود، ناهمگونی محیطی بالا، روابط پیچیده و نسبت سیگنال به نویز پایین قابل انتظار است. بنابراین مقادیر منفی R^2 در این مطالعه به عنوان نشانه‌ای از محدودیت توان تعمیم‌پذیری مدل در برخی زیرنمونه‌ها تفسیر می‌شود و نه به عنوان خطای محاسباتی مدل. کاربرد روش‌های بازنمونه‌گیری و ارزیابی پایداری مدل‌ها در سال‌های اخیر به یکی از رویکردهای متداول در مطالعات یادگیری ماشین محیطی تبدیل شده است. در این پژوهش نیز از بوت‌استرپ برای ارزیابی حساسیت نتایج نسبت به تغییر ترکیب نمونه‌ها و بررسی میزان پایداری عملکرد مدل‌ها استفاده شد.



شکل ۶. ارزیابی مقایسه‌ای مدل‌های جنگل تصادفی، **XGBoost** و **Gradient Boosting** در پیش‌بینی درصد پوشش گیاهی؛ **A-C**: نمودارهای مقادیر مشاهده‌شده در برابر پیش‌بینی‌شده به همراه فاصله اطمینان ۹۵ درصد؛ **D-F**: احتمال پوشش فاصله پیش‌بینی و عرض فاصله پیش‌بینی؛ **G-I**: تجزیه عدم قطعیت به مؤلفه‌های آلانوریک و اپستیمیک؛ **J-L**: اهمیت ویژگی‌ها بر اساس مقادیر **SHAP**؛ **M-O**: سهم ویژگی‌های خاک در عدم قطعیت پیش‌بینی

در نمودارهای پیش‌بینی با فواصل اطمینان (شکل ۶D-F)، دو شاخص کلیدی احتمال پوشش فاصله پیش‌بینی و عرض فاصله پیش‌بینی مورد بررسی قرار گرفتند. Gradient Boosting (شکل ۶F) با PICP برابر با ۶۰ درصد و عرض میانگین PI معادل ۱۸/۷ درصد، بهترین پوشش را ارائه داد، اگرچه هنوز با مقدار ایده‌آل ۹۵ درصد فاصله داشت. در مقابل، XGBoost (شکل ۶E) با PICP برابر ۵۵ درصد و عرض PI باریک‌تر (۱۶/۸ درصد)، نشان داد علی‌رغم دقت پیش‌بینانه ظاهری بالاتر، فواصل اطمینان

محافظه کارانه تری تولید می کند. این تضاد بین دقت نقطه ای و قابلیت اطمینان که در مطالعات پیشرفته نقشه برداری دیجیتال خاک و اکولوژی نیز گزارش شده است، انتخاب مدل را به تصمیمی وابسته به زمینه کاربردی تبدیل می کند (وس و لاگاشیری^۱، ۲۰۱۷).

تجزیه کمی منابع عدم قطعیت و تحلیل اهمیت ویژگی ها

تجزیه عدم قطعیت به دو جزء آلتوری (ذاتی) و ایپستمیک (مدلی)، منشأ اصلی خطاها را مشخص کرد (شکل ۶G-I). در هر سه مدل، سهم عدم قطعیت ایپستمیک به وضوح غالب بود (بیش از ۵۰ درصد در هر مدل). این یافته حائز اهمیت است زیرا نشان می دهد خطای اصلی از محدودیت های دانش مدل (نظیر تعاملات ناشناخته، متغیرهای حذف شده یا ساختار الگوریتمی ناکافی) نشئت می گیرد. غلبه عدم قطعیت ایپستمیک، مطابق با یافته های دیتسه (۲۰۱۷)، نشان می دهد بهبود پارادایم مدل سازی و گنجاندن متغیرهای فرایندمحور ضروری تر از صرفاً افزایش دقت اندازه گیری هاست.

بحث

تحلیل SHAP که مبتنی بر تئوری بازی ها است، معیار مدل آگنوستیک و بی طرفانه برای تفسیر اهمیت ویژگی ها ارائه داد. در این مطالعه، با توجه به عملکرد برتر مدل XGBoost (میان $R^2 = 0.50$ در داده های ترکیبی و 0.76 در منطقه دره)، تحلیل SHAP بر اساس این مدل انجام شد. نتایج نشان داد ماده آلی، شن و TNV به طور ثابتی در میان ویژگی های با بیشترین اهمیت قرار دارند. این یافته با نتایج حاصل از روش های Gini Importance و Permutation Importance (شکل ۴) که ماده آلی را در رتبه دوم اهمیت قرار داده بود، همخوانی نسبی دارد و قابلیت اطمینان این شناسایی را تقویت می کند.

اهمیت بالای ماده آلی نه تنها به عنوان یک منبع غذایی، بلکه به دلیل بهبود ساختار خاک، نفوذپذیری آب، ظرفیت تبادل کاتیونی و فعالیت میکروبی است که می تواند اثرات منفی محتوای بالای رس و آهک (TNV) را تعدیل کند. درصد شن نیز به دلیل نقش آن در زهکشی و تهویه خاک، به عنوان یک عامل فیزیکی مؤثر بر استقرار گیاه شناخته می شود. TNV (ارزش خنثی سازی کل) به عنوان شاخصی از محتوای کربنات کلسیم خاک، نشان دهنده نقش بازدارندگی خاک های آهکی در جذب عناصر غذایی و کاهش دسترسی زیستی فسفر است.

قابل توجه است که فسفر اگرچه در تحلیل منطقه ای (شکل ۳-E) به عنوان مهم ترین عامل محدودکننده در ۷۵ درصد مناطق شناسایی شد، در تحلیل SHAP در رتبه های پایین تری قرار گرفت. این تفاوت نشان دهنده آن است که تأثیر فسفر بر پوشش گیاهی ممکن است بیشتر در مقیاس محلی (منطقه ای) قابل شناسایی باشد تا در سطح مدل سراسری. با این حال، این یافته با مفهوم «تله فسفاتی» در خاک های غنی از کلسیم همخوانی داشته و بر اهمیت مدیریت فسفر در این خاک ها تأکید می کند.

جالب توجه است که رس که در تحلیل Gini Importance (شکل ۴-A) به عنوان مهم ترین متغیر شناسایی شده بود، در تحلیل SHAP در رده های پایین تری قرار گرفت. این مشاهده نشان می دهد تأثیر رس بر پوشش گیاهی در این منطقه، اگرچه دارای نقش غیرخطی و آستانه ای مهمی است (شکل ۳-G)، اما در مدل سراسری XGBoost سهم کمتری در پیش بینی ها دارد. با این حال، نباید از نقش غیرخطی و آستانه ای بافت خاک غافل شد؛ به طوری که نمودار وابستگی جزئی رس (شکل ۳-G) وجود یک آستانه حدود $12/3$ درصد را نشان می دهد. این پدیده - که یک متغیر در مدل های پیشرفته اهمیت پایین تری دارد، ولی در شرایط خاص (آستانه ای) تأثیر قوی نشان می دهد - بارها در مطالعات پدولوژی گزارش می شود و لزوم استفاده از روش های تحلیل چندمتغیره و غیرخطی را تأیید می کند (پادریان و همکاران، ۲۰۲۰).

از آنجا که تمرکز این پژوهش بر مقایسه نسبی عملکرد الگوریتم های یادگیری ماشین و شناسایی عوامل کنترل کننده پوشش گیاهی بود، ارزیابی مدل ها بر پایه چارچوب اعتبارسنجی متعارف و تحلیل عدم قطعیت مبتنی بر بوت استرپ انجام شد. با این حال، استفاده از رویکردهای اعتبارسنجی مکانی نظیر Spatial Cross-Validation و بررسی خودهمبستگی مکانی باقیمانده ها از طریق

شاخص Moran's I می‌تواند در مطالعات آینده به درک بهتر ساختار فضایی خطاها و ارزیابی محافظه کارانه‌تر قابلیت تعمیم مدل‌ها کمک کند.

نتیجه‌گیری و پیشنهادها

این مطالعه با هدف تحلیل روابط پیچیده خاک و پوشش گیاهی در منطقه ناهمگن آبخیز زوجی شوش انجام شد. یافته‌های اصلی نشان می‌دهد روابط خاک - گیاه در منطقه مورد مطالعه غیرخطی، آستانه‌ای و به شدت وابسته به ناهمگونی توپوگرافی است. مدل‌سازی تفکیکی (جداگانه برای تپه و دره) نسبت به مدل یکپارچه برتری چشمگیری داشت؛ به طوری که مدل‌های مبتنی بر boosting (XGBoost و Gradient Boosting) در زیرمجموعه همگن دره عملکرد بسیار خوبی نشان دادند ($R^2 \approx 0.75$)، در حالی که در داده‌های ترکیبی، Random Forest به دلیل مقاومت بیشتر در برابر ناهمگونی، عملکرد پایدارتری ارائه داد. این یافته نشان می‌دهد که انتخاب الگوریتم باید بر اساس ویژگی‌های داده و میزان ناهمگونی محیطی انجام شود.

در سطح ویژگی‌های خاک، فسفر به عنوان عامل محدودکننده اصلی در ۷۵ درصد مناطق شناسایی شد. نتایج مدل‌های پیش‌بینی نشان داد فسفر حتی در شرایط pH نسبتاً قابل تحمل، همچنان محدودیت اصلی در دسترسی گیاهان به مواد مغذی است که تأکید بر اهمیت اصلاح چرخه فسفر در خاک‌های آهکی دارد. همچنین، نتایج همبستگی پیرسون نشان داد روابط خطی به‌تنهایی قادر به تبیین تغییرات پوشش گیاهی نیستند و استفاده از مدل‌های یادگیری ماشین برای آشکارسازی روابط غیرخطی و چندمتغیره ضروری است. بر اساس تحلیل SHAP (به عنوان مبنای نتیجه‌گیری نهایی)، ماده آلی، شن و TNV مهم‌ترین عوامل کنترل‌کننده پوشش گیاهی هستند. علاوه بر این، با وجود همبستگی خطی ضعیف رس با پوشش گیاهی ($r \approx -0.10$)، این متغیر به دلیل رابطه آستانه‌ای با پوشش گیاهی (آستانه حدود ۱۲/۳ درصد) در مدل Gini Importance به عنوان مهم‌ترین پیش‌بین شناسایی شد. این یافته نشان می‌دهد روش‌های خطی کلاسیک قادر به کشف روابط پنهان غیرخطی نیستند و استفاده از مدل‌های یادگیری ماشین ضروری است.

ارزیابی عدم قطعیت نشان داد عدم قطعیت‌های موجود عمدتاً ناشی از محدودیت‌های دانش مدل (عدم قطعیت اپیستمیک) است تا خطاهای اندازه‌گیری. این موضوع تأکید می‌کند که بهبود پارادایم مدل‌سازی و گنجانیدن متغیرهای فرایندمحور (مانند رطوبت خاک و فعالیت میکروبی) می‌تواند مؤثرتر از صرفاً افزایش دقت اندازه‌گیری‌ها باشد. این یافته‌ها به‌ویژه در زمینه‌های مدیریتی و برنامه‌ریزی برای احیای پوشش گیاهی در خاک‌های محدودکننده کاربرد دارند.

سپاسگزاری

به این‌وسیله از حمایت‌های مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی خوزستان جهت انجام این طرح پژوهشی سپاسگزاری به عمل می‌آید.

تضاد منافع نویسندگان

نویسندگان این مقاله اعلام می‌دارند که هیچ‌گونه تضاد منافی در خصوص نگارش و انتشار مطالب و نتایج این پژوهش ندارند.

دسترسی به داده‌ها

داده‌ها و نتایج استفاده‌شده در این پژوهش از طریق مکاتبه با نویسنده مسئول در اختیار قرار خواهد گرفت.

مشارکت نویسندگان

نویسنده اول: تحلیل رسمی و بررسی آماری داده‌ها، تهیه و ترسیم نمودارها و داده‌های تصویری

نویسنده دوم: مفهوم‌سازی، تدوین روش‌شناسی

نویسنده سوم: ویرایش و بازبینی نسخه نهایی مقاله

نویسنده چهارم: نگارش پیش‌نویس اولیه مقاله

منابع

- 1 Baltensweiler, A., Walthert, L., Hanewinkel, M., Zimmermann, S., & Nussbaum, M. (2021). Machine learning based soil maps for a wide range of soil properties for the forested area of Switzerland. *Geoderma Regional*, 27, Article e00437. <https://doi.org/10.1016/j.geodrs.2021.e00437>.
- 2 Cao, J., Zhang, Z., Wang, C., Liu, J., & Zhang, L. (2019). Susceptibility assessment of landslides triggered by earthquakes in the Western Sichuan Plateau. *Catena*, 175, 63-76.
- 3 Dietze, M. C. (2017). Ecological forecasting. *Princeton University Press*. doi: 10.22098/mmws.2025.17919.1633.
- 4 Food and Agriculture Organization of the United Nations. (2021). The state of the world's land and water resources for food and agriculture: Systems at breaking point (SOLAW 2021). *FAO*. <https://doi.org/10.4060/cb9910en>
- 5 Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression.
- 6 Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12. <https://doi.org/10.1021/ci0342472>
- 7 Heidary, K., Najafinejad, A. , Mohammadian Behbahani, A. , Ownegh, M. and W. Dekker, L. (2025). Application of Machine Learning Algorithms in Analyzing the Effect of Fire on Soil Water Repellency in Tushan Watershed, Golestan Province. *Watershed Management Research*, (In Persian). doi: 10.22092/wmrj.2025.370992.1637
- 8 Heidary, K. , Najafinejad, A. , Mohammadian Behbahani, A. , Ownegh, M. and W. Dekker, L. (2026). Application of Machine Learning Algorithms in Analyzing the Effect of Fire on Soil Water Repellency in Tushan Watershed, Golestan Province. *Watershed Management Research*, 39(2), 20-43. doi: 10.22092/wmrj.2025.370992.1637
- 9 Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., ... & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, 12(2), Article e0169748. <https://doi.org/10.1371/journal.pone.0169748>.
- 10 Hinsinger, P. (2001). Bioavailability of soil inorganic P in the rhizosphere as affected by root-induced chemical changes: A review. *Plant and Soil*, 237(2), 173–195. <https://doi.org/10.1023/A:1013351617532>
- 11 Jenny, H. (1941). Factors of soil formation: A system of quantitative pedology. *McGraw-Hill*.
- 12 Kopittke, P. M., Menzies, N. W., Wang, P., McKenna, B. A., & Lombi, E. (2019). Soil and the intensification of agriculture for global food security. *Environment International*, 132, Article 105078. <https://doi.org/10.1016/j.envint.2019.105078>

- 13 Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774) .
- 14 McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1–2), 3–52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- 15 Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. *Ecological Modelling*, 411, Article 108815. <https://doi.org/10.1016/j.ecolmodel.2019.108815>
- 16 Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, 264, 301–311. <https://doi.org/10.1016/j.geoderma.2015.07.017>
- 17 Mohammadian Behbahani, A. , Heidary, K. and Hosseinalizadeh, M. (2025). Modeling soil water repellency in loess soils of northern Iran using machine learning. *Water and Soil Management and Modelling*. (In Persian)
- 18 Mousavi, S. R. , Parsayi, F. , Rahmani, A. , Sedri, M. H. & kohsar Bostani, M. (2020). Spatial Prediction Some of the Surface Soil Properties Using Interpolation and Machine Learning Models. *Journal of Soil Management and Sustainable Production*, 10(3), 27-49. doi: 10.22069/ejsms.2021.17251.1916. (In Persian)
- 19 Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *Soil*, 6(1), 35-52.
- 20 Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406, 109–120. <https://doi.org/10.1016/j.ecolmodel.2019.06.002>.
- 21 Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, 11, 1120. <https://doi.org/10.3389/fpls.2020.01120>
- 22 Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, 291, 55–64.
- 23 Wadoux, A. M. J. C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210, Article 103359. <https://doi.org/10.1016/j.earscirev.2020.103359>.
- 24 Winzeler, H. E., Mancini, M., Blackstock, J. M., Libohova, Z., Owens, P. R., Ashworth, A. J., Silva, S. H. (2024). Vegetation masking of remote sensing data aids machine learning for soil fertility prediction. *Remote Sensing*, 16(17), Article 3297. <https://doi.org/10.3390/rs16173297>
- 25 Yang, R. M., Zhang, G. L., Liu, F., Lu, Y. Y., Yang, F., Yang, F., ... & Li, D. C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecological indicators*, 60, 870-878.
- 26 Yu, X., Liu, Y., Niu, S., Zhao, W., Fu, C., & Chen, Z. (2021). Structure, functions, and interactions of dryland ecosystems. In X. Yu & Y. Liu (Eds.), *Dryland social-ecological systems in changing environments* (pp. 69–107). *Springer*. https://doi.org/10.1007/978-981-16-3128-3_3

- 27 Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., & Finke, P. (2019). Digital mapping of soil properties using multiple machine learning models in the semi-arid region of central Iran. *Geoderma*, 337, 159-171. <https://doi.org/10.1016/j.geoderma.2018.09.006>
- 28 Zhang, L., Zhang, Z., Luo, Y., Cao, J., & Tao, F. (2021). Combining optical, fluorescence, thermal satellite, and environmental data to predict county-level maize yield in China using machine learning. *Remote Sensing of Environment*, 253, 112185. <https://doi.org/10.3390/rs12010021>.
- 29 Zornoza, R., Acosta, J. A., Bastida, F., Domínguez, S. G., Toledo, D. M., & Faz, A. (2015). Identification of sensitive indicators to assess the interrelationship between soil quality, management practices and human health. *Soil*, 1(1), 173–185. <https://doi.org/10.5194/soil-1-173-2015>.