

Original Article



The Effectiveness of Data Balancing Approaches in Digital Soil Mapping (Case Study: A Part of Zanjan Province Lands)

Received: 2024.08.31

Accepted: 2024.09.30

**Mastane Rahimi Mashkaleh,¹ Mohammad Amir Delavar,^{1*} 
Mohammad Jamshidi²**

¹ Department of Soil Science, Faculty of Agriculture, University of Zanjan, Iran

² Institute of Soil and Water Research, Agricultural Research, Education and Extension Organization, Karaj, Iran

EXTENDED ABSTRACT

Introduction: Digital soil mapping using innovative machine learning methods is increasingly used to predict the spatial distribution and various soil properties. However in soil science studies, the use of digital soil mapping methods faces challenges due to the imbalance in soil classes, which negatively affects the performance of machine learning algorithms. Therefore, this study aims to address this challenge by improving the classification of imbalanced soil classes through two approaches: resampling and cost-sensitive learning, using the random forest prediction model in Zanjan Province.

Material and Methods: A number of 148 soil samples were collected based on a random classification pattern with a 500 meter spacing and subjected to various physical and chemical analyses in the laboratory following standard methods. Environmental covariates included geomorphological and geological maps, digital elevation model (DEM), and Landsat 8 satellite images, which were selected as inputs for soil class prediction based on expert opinion and principal component analysis (PCA). Some environmental covariates, such as geomorphological and geological maps information and features extracted from DEM, were identified as the most effective predictors for soil classes and were chosen as model inputs. Analytical hill shading (AHS), sunrise, valley depth, LS_factor, channel network distance (CND), topographic wetness index (TWI) and multi-resolution ridge top flatness index (MRRTF) were selected as the most effective environmental variables and modeled the most spatial variability of the soils of the region. Soil-landscape relationship modeling was done performed using Random Forest algorithm and correcting imbalanced data was done by resampling approach using `ubOver` and `ubUnder` functions and also by cost-sensitive learning approach using `rf` function in Random Forest package in Rstudio software environment.

Results and discussion: Soil subgroups were classified into five imbalanced classes, including Typic Calcixerepts, Typic Haploxerepts, GypsicHaploxerepts, Typic Xerorthents, and Lithic Xerorthents. The validation results showed that the overall accuracy (OA) and kappa coefficient for evaluating the soil map with imbalanced data were 65% and 0.32, respectively. After data balancing through resampling, these values increased to 71% and 0.54, respectively, and in the cost-sensitive learning approach, they reached 86% and 0.77, respectively. GypsicHaploxerepts and Lithic Xerorthents subgroups, considered minority classes, were unidentified and excluded when using imbalanced classes. However, after data improvement and augmentation with both resampling and cost-sensitive learning approaches, the prediction of these two minority classes demonstrated acceptable accuracy improvements.

Conclusion: The results of the evaluation of the models showed that in modeling using an unbalanced distribution of soil classes, due to the loss of classes with a small number of observations, uncertain maps with relatively poor accuracy are created, and after applying data balancing, the accuracy of models based on soil relationships - Topography is improved in digital soil mapping studies. The results showed that the cost-sensitive learning approach, focusing on classes with low repetition, can be used as a superior model in other areas. Considering that the research in the field of unbalanced soil data is limited, this study can be an effective solution to deal with unbalanced data in soil classes and produce digital soil maps with high accuracy.

Key words: Random forest, Imbalanced data, Resampling, Cost-sensitive learning.

How to cite this article:
Rahimi Mashkaleh, M.,
Delavar, M.A. and Jamshidi,
M., 2025. The Effectiveness
of Data Balancing
Approaches in Digital Soil
Mapping (Case Study: A
Part of Zanjan Province
Lands). Adv. Environ. Sci.
23(3): 655-670.

* Corresponding Author Email Address: amir-delavar@znu.ac.ir

DOI: 10.48308/envs.2024.1433



Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license <https://creativecommons.org/licenses/by/4.0/>.

کارایی رویکردهای متعادل سازی داده در نقشه برداری رقومی خاک (مطالعه موردی: بخشی از اراضی استان زنجان)

تاریخ دریافت: ۱۴۰۳/۰۶/۱۰

تاریخ پذیرش: ۱۴۰۳/۰۷/۰۹

مستانه رحیمی مشکله^۱، محمد امیر دلاور^{۲*}، محمد جمشیدی^۳

^۱ دانش آموخته دکتری گروه علوم خاک دانشکده کشاورزی دانشگاه زنجان

^۲ گروه علوم خاک دانشکده کشاورزی، دانشگاه زنجان

^۳ موسسه تحقیقات خاک و آب، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج

چکیده مبسوط

سابقه و هدف: نقشه برداری رقومی خاک با استفاده از روش های نوین یادگیری ماشین به طور گسترده ای برای پیش بینی پراکندگی مکانی ویژگی های مختلف خاک به کار گرفته می شود، با این وجود یکی از محدودیت های استفاده از روش های نقشه برداری رقومی خاک در مطالعات خاکشناسی، عدم تعادل کلاس های خاک است که تأثیر منفی بر عملکرد الگوریتم های یادگیری ماشین دارد؛ بنابراین این پژوهش برای رفع این چالش و بهبود طبقه بندی کلاس های نامتعادل خاک با دو رویکرد نمونه گیری مجدد و یادگیری حساس به هزینه و استفاده از مدل پیش بینی جنگل تصادفی در استان زنجان انجام گرفته است.

مواد و روش ها: تعداد ۱۴۸ خاک رخ مشاهداتی بر اساس الگوی طبقه بندی تصادفی با فاصله ۵۰۰ متر حفر و پس از انتقال به آزمایشگاه تجزیه های مختلف فیزیکی و شیمیایی مطابق با روش های استاندارد بر روی آن ها انجام گرفت. متغیرهای محیطی شامل اطلاعات نقشه های ژئومورفولوژی و زمین شناسی، مدل رقومی ارتفاع و داده های حاصل از تصاویر ماهواره ای لندست ۸ بودند که بر اساس نظر کارشناسی و رویکرد تحلیل مؤلفه اصلی تعدادی از متغیرهای محیطی شامل اطلاعات نقشه های ژئومورفولوژی، اطلاعات زمین شناسی و ویژگی های مستخرج از مدل رقومی ارتفاع به عنوان مؤثرترین متغیرهای محیطی برای پیش بینی کلاس های خاک و به عنوان ورودی مدل انتخاب شدند. مدل سازی رابطه خاک-زمین نما با استفاده از الگوریتم جنگل تصادفی و اصلاح داده های نامتعادل توسط رویکرد نمونه گیری مجدد با استفاده از توابع ubOver و ubUnder و همچنین رویکرد یادگیری حساس به هزینه با استفاده از تابع rf در بسته Random Forest در محیط برنامه نویسی Rstudio انجام شد.

نتایج و بحث: نتایج این پژوهش حاکی از این بود که خاک های منطقه در سطح زیرگروه در پنج کلاس با توزیع نامتعادل شامل تیپیک کلسی زریپتز، تیپیک هاپلوزریپتز، جیپسیک هاپلوزریپتز، تیپیک زراورتنز و لیپیک زراورتنز طبقه بندی شدند. بر این اساس مقادیر آماره های صحت کلی و ضریب کاپا برای ارزیابی نقشه خاک با داده های نامتعادل به ترتیب برابر ۶۵ درصد و ۰/۳۲ بوده و پس از متعادل سازی داده ها در رویکرد نمونه گیری مجدد به ترتیب برابر ۷۱ درصد و ۰/۵۴ و در رویکرد یادگیری حساس به هزینه به ترتیب برابر ۸۶ درصد و ۰/۷۷ به دست آمد. خاک های زیرگروه های جیپسیک هاپلوزریپتز و لیپیک زراورتنز که جزء کلاس های اقلیت محسوب می شدند، هنگام استفاده از کلاس های نامتعادل پیش بینی نشده و حذف شده بودند اما پس از بهبود داده ها و بیش افزایی با دو رویکرد نمونه گیری مجدد و یادگیری حساس به هزینه به تعداد این دو کلاس اقلیت، پیش بینی این زیرگروه ها با صحت قابل قبولی افزایش نشان داد.

نتیجه گیری: نتایج ارزیابی مدل ها نشان داد که در مدل سازی با استفاده از توزیع نامتعادل کلاس های خاک، به دلیل از دست رفتن کلاس های با تعداد مشاهده کم، نقشه های نامطمئن با دقت نسبتاً ضعیفی ایجاد می شود که پس از اعمال متعادل سازی داده ها، دقت مدل های مبتنی بر روابط خاک - زمین نما در مطالعات نقشه برداری رقومی خاک ارتقا می یابد. نتایج نشان داد که رویکرد یادگیری حساس به هزینه با تمرکز بر روی کلاس های با تکرار کم، می تواند به عنوان یک مدل برتر در مناطق دیگر نیز مورد استفاده قرار گیرد. با توجه به اینکه تحقیقات در زمینه داده های نامتعادل در خاک محدود است، این مطالعه می تواند یک راه حل مؤثر برای مقابله با داده های نامتعادل در کلاس های خاک و تولید نقشه های رقومی خاک با دقت بالا باشد.

واژه های کلیدی: جنگل تصادفی، داده های نامتعادل، نمونه گیری مجدد، یادگیری حساس به هزینه.

استناد به این مقاله: رحیمی مشکله، م.، دلاور و م.، جمشیدی. ۱۴۰۴. کارایی رویکردهای متعادل سازی داده در نقشه برداری رقومی خاک (مطالعه موردی: بخشی از اراضی استان زنجان). فصلنامه علوم محیطی نوین. ۶۷۰: ۶۵۵-۶۵۵(۳).

* Corresponding Author Email Address: amir-delavar@znu.ac.ir

DOI: 10.48308/envs.2024.1433



مقدمه

از منطقه مورد مطالعه با تعداد مشاهده‌های کم‌تر شود. در واقع مشکل عدم تعادل در داده‌ها باعث کاهش دقت و از دست رفتن کلاس‌های اقلیت (کم تعداد) در پیش‌بینی‌ها و ایجاد نقشه‌های نامطمئن یا گمراه‌کننده می‌شود (Sharififar *et al.*, 2019b).

رویکردهای متفاوتی برای غلبه بر مشکل عدم تعادل کلاس‌های خاک وجود دارد اما ازجمله رویکردهای یادگیری ماشین برای پرداختن به این مسئله که توسط محققان علوم خاک ارائه شده است می‌توان نمونه‌گیری مجدد^۱ و یادگیری حساس به هزینه^۲ را نام برد (Sharififar *et al.*, 2019a; Taghizadeh-Mehrjardi *et al.*, 2019; Rahimi *et al.*, 2023a). روش نمونه‌گیری مجدد می‌تواند عدم تعادل داده‌های آموزشی را با تعدیل تعداد نمونه‌ها از کلاس‌های اکثریت و اقلیت، داده‌های آموزشی متعادل کند (Krawczyk, 2016). رویکرد یادگیری حساس به هزینه، با به حداقل رساندن هزینه طبقه‌بندی اشتباه موجب بهبود دقت نقشه تولیدی شده و عملکرد بهتری در توزیع کلاس‌های نامتعادل دارد (Vincent *et al.*, 2019b; Sharififar *et al.*, 2018).

باوجود اینکه نقشه‌برداری رقومی خاک به‌عنوان یک استراتژی مؤثر برای افزایش بهره‌وری خاک در تولید پایدار مورد استفاده قرار می‌گیرد. اما نامتوازن بودن داده‌ها در این مطالعات به‌عنوان یک چالش در نظر گرفته می‌شود؛ بنابراین این پژوهش باهدف دستیابی به تحلیل دقیق‌تر و کاربردی‌تر داده‌های خاک و تولید نقشه‌های با دقت بالاتر به بررسی چالش‌ها و راه‌حل‌های مرتبط با پیش‌بینی مکانی کلاس‌های نامتعادل خاک پرداخته و به‌منظور بهبود دقت طبقه‌بندی کلاس‌های نامتعادل خاک، عملکرد رویکردهای مختلف متعادل‌سازی از قبیل نمونه‌گیری مجدد از داده‌ها و یادگیری حساس به هزینه با استفاده از مدل یادگیری ماشین جنگل تصادفی را در بخشی از اراضی استان زنجان مورد مقایسه و ارزیابی قرار می‌دهد.

نقشه‌برداری رقومی خاک با استفاده از فن‌های نوآورانه یادگیری ماشین به‌طور فزاینده‌ای در پیش‌بینی توزیع مکانی و ویژگی‌های مختلف خاک به کار می‌رود (Brungard *et al.*, 2015; Khaledian and Miller, 2015; Heung *et al.*, 2016). نقشه‌برداری رقومی خاک به‌عنوان یک چارچوب نمایش مکانی خاک با توجه به نتایج کمی، تکرارپذیری و قابلیت تحلیل عدم اطمینان، جایگاه خود را تثبیت کرده است (Fantappiè *et al.*, 2023)؛ اما علی‌رغم افزایش دقت روش‌های نقشه‌برداری رقومی خاک در سال‌های اخیر، تولید نقشه‌های خاک در مقیاس منطقه‌ای با دقت بالا همچنان یک فعالیت چالش‌برانگیز است (Meng *et al.*, 2022).

در چند دهه گذشته، رویکردهای نقشه‌برداری رقومی خاک توسعه یافته‌اند تا اطلاعات و استنباط‌های خاک را به‌صورت پیوسته ارائه دهند (Minasny *et al.*, 2013; Malone *et al.*, 2003; McBratney *et al.*, 2009; *al.*). ظهور و بهبود مداوم روش‌های یادگیری ماشین و رویکردهای سنجش‌ازدور منجر به پیش‌بینی موفق ویژگی‌های خاک با استفاده از فن‌های نقشه‌برداری رقومی خاک شده است که می‌تواند به حل محدودیت‌های روش‌های مرسوم نقشه‌برداری خاک کمک کند (Minasny and Hartemink, 2011; Wadoux *et al.*, 2020).

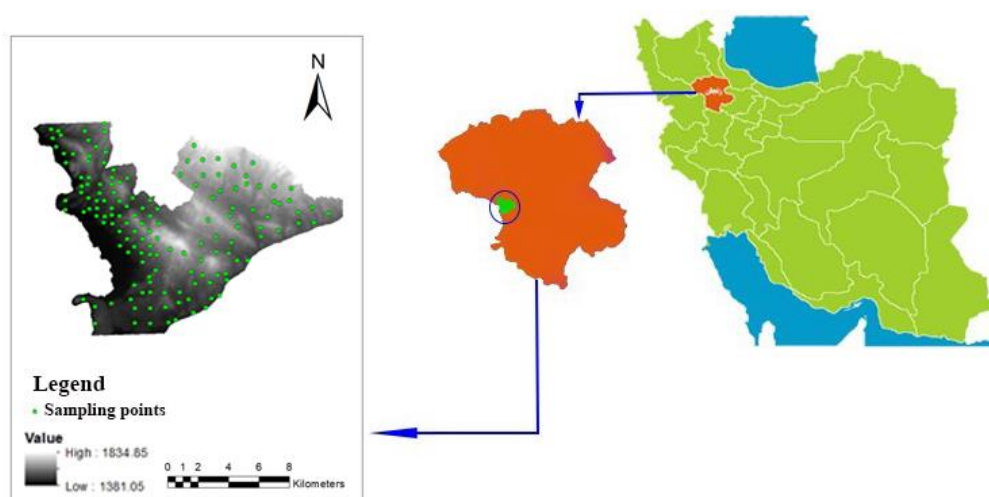
یکی از مسائل و مشکلات شایع که خاک‌شناسان و پژوهشگران عمدتاً با آن مواجه می‌شوند، عدم تعادل در تعداد مشاهده‌ها برای انواع مختلف خاک است. این عدم تعادل ممکن است ناشی از وجود عواملی باشد که در فرآیند تشکیل و تکامل خاک تأثیر دارند (Heung *et al.*, 2016; Taghizadeh-Mehrjardi *et al.*, 2020). تعداد نامتعادل داده‌ها در کلاس‌های مشاهده‌شده خاک یک منطقه ممکن است منجر به برآورد ناکافی کلاس‌های اقلیت در مدل‌سازی و تخمین بیش‌ازحد کلاس‌های اکثریت در مدل‌سازی شود. به‌عبارت دیگر، این پدیده ممکن است باعث حذف یک ناحیه

مواد و روش ها

منطقه مورد مطالعه با مساحتی حدود ۱۳۸۲۳ هکتار، بخشی از اراضی جنوب غربی استان زنجان است که در مختصات جغرافیائی ۴۷ درجه و ۲۱ دقیقه تا ۴۸ درجه و ۱۱ دقیقه طول شرقی و ۳۶ درجه و ۳۷ دقیقه تا ۳۶ درجه و ۳۱ دقیقه عرض شمالی واقع شده است (شکل ۱). میانگین ارتفاع منطقه مورد مطالعه از سطح دریا ۱۴۸۲ متر است. با توجه به آمار بلندمدت ۲۰ ساله، متوسط بارندگی سالانه منطقه ۳۴۰ میلی متر و به ترتیب متوسط دمای سالیانه ۱۳، بیشینه دمایی ۱۴/۸۱ و کمینه دمایی ۴/۸۵ درجه سلسیوس است

(Statistical Yearbook of Zanjan Province, 2019).

فیزیوگرافی منطقه شامل دو واحد اراضی تپه ماهور^۳ و دشت های دامنه ای^۴ و عمده سازندهای زمین شناسی منطقه شامل لایه های کربناته، سنگ آهک، کنگلومرا و مواد آتشفشانی و آذرآواری مانند توف اسیدی است. منطقه دارای پوشش گیاهی تنک است و در زمره مراتع ضعیف قرار می گیرد. با استناد به نقشه رژیم های رطوبتی و حرارتی ایران، منطقه مذکور دارای رژیم حرارتی مزیک^۵ و رژیم رطوبتی زیریک^۶ هستند (Soil and Water Research Institute, 2010).



شکل ۱- موقعیت منطقه مورد مطالعه و نقاط نمونه برداری
Fig. 1- Location of the study area and sampling points

عملیات صحرایی و تجزیه های آزمایشگاهی

مطابق روش های استاندارد مطالعات خاک شناسی تعداد ۱۴۸ خاک رخ بر اساس الگوی طبقه بندی تصادفی با میانگین فاصله ۵۰۰ متر (در برخی نقاط بر اساس نظر کارشناس و شرایط محلی تا فاصله ۷۰۰ متر تغییر یافتند)، حفر شدند (Soil science division staff, 2017). نمونه برداری و تشریح تمامی خاک رخی های حفر شده بر اساس راهنمای تشریح خاک رخی ها انجام گردید (Schoeneberger et al., 2012). نمونه ها پس از هوا خشک شدن از الک دو میلی متری عبور داده شدند و تجزیه های فیزیکی و شیمیایی با استفاده از روش های استاندارد و متداول انجام گرفت

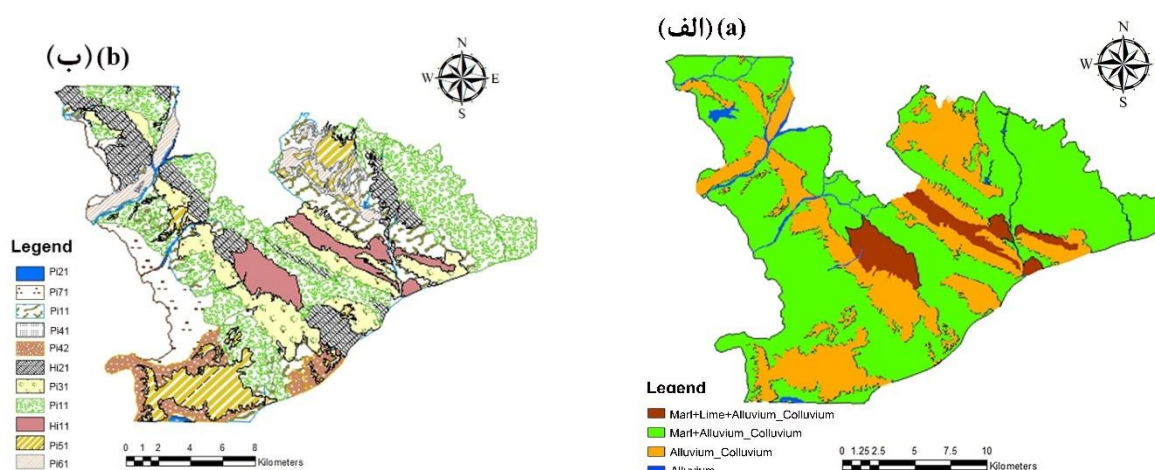
(USDA, 2004). خاک رخی ها بر اساس سیستم جامع رده بندی خاک به روش آمریکایی تا سطح فامیل طبقه بندی شدند (Soil Survey Staff, 2022).

استخراج متغیرهای محیطی

اطلاعات نقشه های ژئومورفولوژی، نقشه زمین شناسی، داده های سنجش از دور و مدل رقومی ارتفاع با قدرت تفکیک مکانی ۳۰ متر (سنجده استر) برای استخراج لایه های کمکی خاک استفاده شد. نقشه ژئومورفولوژی منطقه بر اساس تلفیق لایه های اطلاعاتی شامل واحدهای لندفرم و مواد مادری به همراه تفسیر تصاویر ماهواره ای بر اساس رویکرد سلسله مراتبی ارائه شده توسط زینک^۷ (۲۰۱۵) تهیه

اتمسفیری در محیط نرم افزار ENVI (نسخه ۵/۳) و ۱۸ متغیر پستی و بلندی از مدل رقومی ارتفاع در محیط نرم افزار SAGA GIS (نسخه ۷/۹) استخراج شد. انتخاب متغیرهای کمکی در این مطالعه بر اساس رویکرد تحلیل مؤلفه اصلی^۸ در نرم افزار SPSS (نسخه ۲۶) و رتبه بندی اهمیت نسبی مدل یادگیری ماشین به همراه نظارت کارشناس انجام شد (Kuhn and Johnson, 2013).

گردید (جدول ۱، شکل ۲-الف). نقشه زمین شناسی با مقیاس ۱:۲۵۰۰۰۰ منطقه تهیه شده توسط سازمان زمین شناسی کشور در محیط نرم افزار Arc-GIS (نسخه ۱۰/۷) زمین مرجع و رقومی شد (شکل ۲-ب). ۳۶ متغیر سنجش از دوری با استفاده از تصاویر سنجنده (OLI/TIRS) ماهواره لندست ۸ با قدرت تفکیک مکانی ۳۰×۳۰ متر (USGS 2014) پس از اعمال تصحیحات رادیومتریکی و



شکل ۲- الف) نقشه زمین شناسی ب) نقشه ژئومورفولوژی منطقه مورد مطالعه

Figure 2- Map of a) geology b) geomorphology in study area

جدول ۱- واحدهای تفکیک شده در سطح لندفرم بر اساس اطلاعات ژئومورفولوژی در منطقه مورد مطالعه

Table 1. Separated units at the landform level based on geomorphological information in the study area

واحد نقشه Map Unit	اجزای لندفرم Landform Components	شکل زمین Landform Shape	سنگ شناسی/منشأ Lithology/Origin	پستی و بلندی/قالب Relief/Morphology	زمین نما Physiography
Consociation	Hi111	Slope facet complex	Marl + Lime + Alluvium - Colluvium	Medium hill	تپه ماهور Hill lands
Complex	Hi211	Summit	Marl + Alluvium - Colluvium	Low hill	
	Hi212	Shoulder			
	Hi213	Backslope			
	Hi214	Footslope			
	Hi215	Toeslope			
Association	Pi111	High glacic	Alluvium - Colluvium	Glacic, Dissected	دشت دامنه‌ای Piedmont plains
Complex	Pi211	Middle glacic	Marl + Alluvium - Colluvium	Glacic, Moderately dissected	
	Pi311	Low glacic	Marl + Alluvium - Colluvium	Glacic, Low dissected	
Association	Pi411	Side slope	Alluvium	Glacic terrace, Dissected	
Consociation	Pi421	Tread	Alluvium - Colluvium		
Association	Pi511	Side slope	Marl + Alluvium - Colluvium	Glacic terrace, Slightly eroded	
Association	Pi611	Side slope	Alluvium - Colluvium	Coalescing fan	
Association	Pi711	Upper part	Marl + Alluvium - Colluvium	Channeled recent alluvial deposits	
	Pi712	Middle part			
	Pi713	Lower part			

مدل‌سازی خاک-زمین‌نما

مدل جنگل تصادفی: مدل جنگل تصادفی یک روش ناپارامتری است که اولین بار توسط بریمن^۹ و همکاران (۱۹۸۴) ارائه شد. این مدل قادر به پیش‌بینی متغیرهای کمی یا متغیرهای طبقه‌بندی‌شده بر اساس مجموعه‌ای از متغیرهای پیش‌بینی‌کننده کمی و کیفی است. در این روش داده‌ها به‌طور تکراری برای به دست آوردن ارتباط بین متغیر پاسخ و متغیرهای مستقل و انجام تخمین جداسازی می‌شوند. در روش جنگل تصادفی برخلاف سایر روش‌های درختی که تعداد محدودی درخت ترسیم می‌کنند، صدها یا هزاران درخت طبقه‌بندی تولید می‌شود (Breiman and Cutler, 2004). این روش یک روش یادگیری گروهی است و برای طبقه‌بندی با ساختن تعداد درختان زیاد عمل می‌نماید (Breiman, 2001). کلیه مراحل مدل‌سازی با استفاده از روش یادگیری جنگل تصادفی با استفاده از بسته Random Forest به همراه کد نویسی در محیط نرم‌افزار RStudio انجام شد.

متعادل‌سازی داده با استفاده از رویکرد نمونه‌گیری مجدد از داده‌ها: در این رویکرد برای اصلاح داده‌های نامتعادل در دو مرحله پیش‌نمونه‌گیری از کلاس‌های خاک با تعداد کمتر (کلاس‌های اقلیت) با استفاده از تابع ubOver و کم‌نمونه‌گیری از کلاس‌های خاک با تعداد بیش‌تر (داده‌های اکثریت) با استفاده از تابع ubUnder در نرم‌افزار Rstudio استفاده شد. مقدار نمونه‌برداری برای کلاس‌های اقلیت تقریباً دو تا سه برابر مقدار اولیه افزایش یافت تا توزیع داده‌ها به یک حد معمول نزدیک شود. قابل‌ذکر است که در هر دو حالت اصلاح داده‌های نامتعادل تلاش گردید تا نسبت اصلی کلاس‌ها تغییر نکنند. این روش‌ها باعث می‌شوند که توزیع داده‌های کلاس‌های خاک متعادل شود (Abdi and Hashemi, 2015).

متعادل‌سازی داده با استفاده از رویکرد یادگیری حساس به هزینه: در این رویکرد، الگوریتم به دنبال

بهینه‌سازی پیش‌بینی‌های کلاس است تا با کاهش کل هزینه طبقه‌بندی اشتباه، بهترین عملکرد را ارائه دهد. در واقع الگوریتم تلاش می‌کند با کاهش تعداد خطاهای طبقه‌بندی، پیش‌بینی بهتر و دقیق‌تری را ارائه دهد. در این روش برای ایجاد یک مدل که با حساسیت به هزینه تغییر یابد، وزن‌هایی به کلاس‌های مختلف اختصاص داده می‌شود. این وزن‌ها از معکوس توزیع کلاس محاسبه می‌شوند، با این ترتیب که الگوریتم متمرکز می‌شود بهترین شکل را برای کلاس اقلیت پیدا کند (Zhang et al., 2021). این تغییرات در مدل مورد استفاده باعث می‌شود که توزیع کلاس‌ها در داده‌های آموزشی مورد توجه قرار گیرد. به‌طور خاص، هر نمونه در داده‌های آموزشی وزنی می‌یابد که بر اساس معکوس فراوانی کلاس مربوط به آن نمونه محاسبه می‌شود. این اقدام باعث می‌شود که نمونه‌های کم‌رخداد (اقلیت) وزن بالاتری در مدل داشته باشند و تأثیر آن‌ها در پیش‌بینی‌های مدل بیشتر شود. در این مطالعه، از تابع rf در بسته Random Forest در محیط برنامه‌نویسی Rstudio برای پیاده‌سازی رویکرد حساس به هزینه استفاده شد.

اعتبارسنجی اعتبار دقت مدل

به‌منظور ارزیابی دقت مدل مورد استفاده، داده‌ها به‌طور تصادفی به دو بخش داده‌های آموزشی و اعتبارسنجی تقسیم شدند. ۸۰ درصد از داده‌ها (۱۱۸ پروفیل) برای آموزش مدل و ۲۰ درصد دیگر (۳۰ پروفیل) به‌عنوان داده‌های اعتبارسنجی برای ارزیابی استفاده شدند. در ادامه مدل با داده‌های آموزش برازش یافت و سپس برای داده‌های اعتبارسنجی پیش‌بینی انجام شد. کلاس‌های پیش‌بینی‌شده با استفاده از ماتریس خطا به‌صورت درصدی گزارش شدند. پارامترهای استخراج‌شده از ماتریس خطا شامل دقت کلی نقشه^{۱۰}، دقت تولیدکننده^{۱۱}، دقت کاربر^{۱۲} و ضریب کاپا^{۱۳} برای اعتبارسنجی به‌کار گرفته شدند (Congalton, 1991; Jensen, 1996).

نتایج و بحث

نتایج انتخاب متغیرهای محیطی برای پیش‌بینی مکانی کلاس‌های خاک در سطح زیرگروه توسط مدل جنگل تصادفی و بر اساس روش تحلیل مؤلفه اصلی و نظارت کارشناس نشان داد که از میان ۵۷ متغیر محیطی تولیدشده در نهایت تعداد ۱۰ متغیر کمکی شامل اطلاعات نقشه‌های ژئومورفولوژی، اطلاعات زمین‌شناسی، مدل رقومی ارتفاع و ویژگی‌های مستخرج از آن شامل تجزیه و تحلیل سایه‌اندازی تپه‌ها^{۱۴}، طلوع خورشید^{۱۵}، عمق دره^{۱۶}، شاخص طول در جهت شیب^{۱۷}، فاصله تا شبکه آبراهه^{۱۸}، شاخص رطوبتی توپوگرافی^{۱۹} و شاخص همواری بالای پشته با درجه تفکیک بالا^{۲۰} به‌عنوان مؤثرترین متغیرهای محیطی انتخاب شدند که بیشترین میزان تغییرپذیری مکانی خاک‌ها در منطقه را مدل‌سازی کنند. همچنین در روش جنگل تصادفی بر اساس شاخص میانگین حداقل صحت اهمیت متغیرهای کمکی منتخب نشان‌دهنده این است که پارامتر عمق دره مؤثرترین متغیر محیطی در مدل‌سازی مکانی خاک‌ها در منطقه مطالعاتی بوده است (شکل ۳). پراکنش مکانی چهار مورد از مهم‌ترین متغیرهای محیطی در پیش‌بینی کننده کلاس‌های خاک در سطح زیرگروه توسط مدل جنگل تصادفی در شکل ۴ نشان داده شده است. بر اساس نتایج از میان متغیرهای محیطی منتخب چهار متغیر وابسته به توپوگرافی شامل عمق دره، فاصله تا شبکه آبراهه، شاخص همواری بالای پشته با درجه تفکیک بالا و شاخص طول در جهت شیب دارای بیش‌ترین اهمیت هستند که نشان می‌دهد توپوگرافی مهم‌ترین عامل خاک‌سازی در منطقه مورد مطالعه است. (Mousavi et al., 2020) در مطالعه خود گزارش کردند که پارامترهای توپوگرافی به‌عنوان مهم‌ترین پیش‌ران‌های محیطی برای مدل‌سازی کلاس‌های خاک هستند.

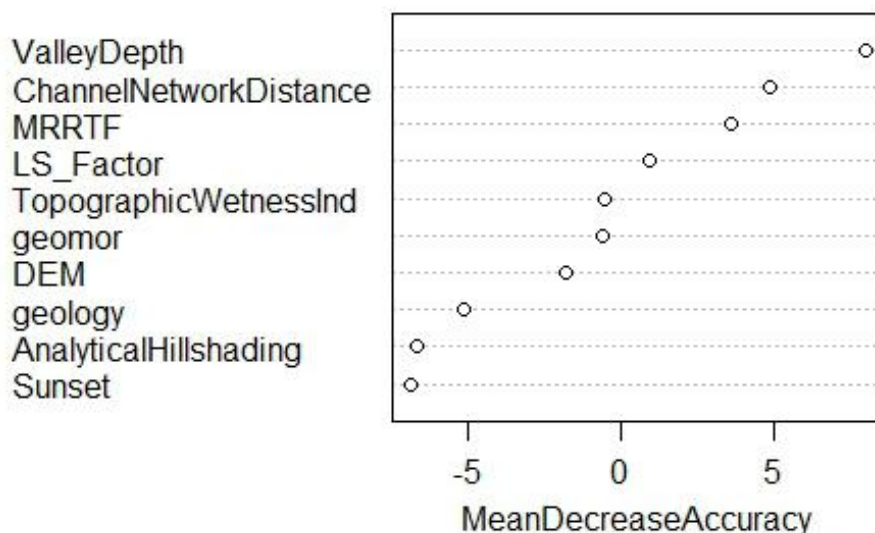
مطابق با نتایج تجزیه ویژگی‌های فیزیکوشیمیایی (جدول ۱) و بر اساس سامانه آمریکایی رده‌بندی خاک‌ها، خاک‌های منطقه در دو رده انتی‌سولز (Entisols) و اینسپتی‌سولز

(Inceptisols) و در سطح زیرگروه در پنج کلاس تیپیک کلسی‌زرپتز (Typic Calcixerepts)، تیپیک هاپلوزرپتز (Typic Haploxerepts)، جیپسیک هاپلوزرپتز (GypsicHaploxerepts)، تیپیک زراورتنتر (Typic Xerorthents) و لیتیک زراورتنتر (Lithic Xerorthents) طبقه‌بندی شدند (Rahimi et al., 2023a). نتایج پیش‌بینی مکانی کلاس‌های خاک منطقه در جدول ۲ نشان داده شده است. نتایج این جدول نشان می‌دهد که زیرگروه‌های جیپسیک هاپلوزرپتز (کلاس C) و لیتیک زراورتنتر (کلاس E) به ترتیب با فراوانی ۸/۱ و ۷/۳۴ درصد به‌عنوان کلاس‌های کم تعداد (اقلیت) و زیرگروه‌های تیپیک کلسی‌زرپتز (کلاس A)، تیپیک هاپلوزرپتز (کلاس B)، تیپیک زراورتنتر (کلاس D) به ترتیب با فراوانی بیشتر از ۳۲/۴۳، ۱۷/۵۶ و ۲۰/۹۴ درصد از کل مشاهده‌های منطقه به‌عنوان کلاس‌های پُر تعداد (اکثریت) هستند.

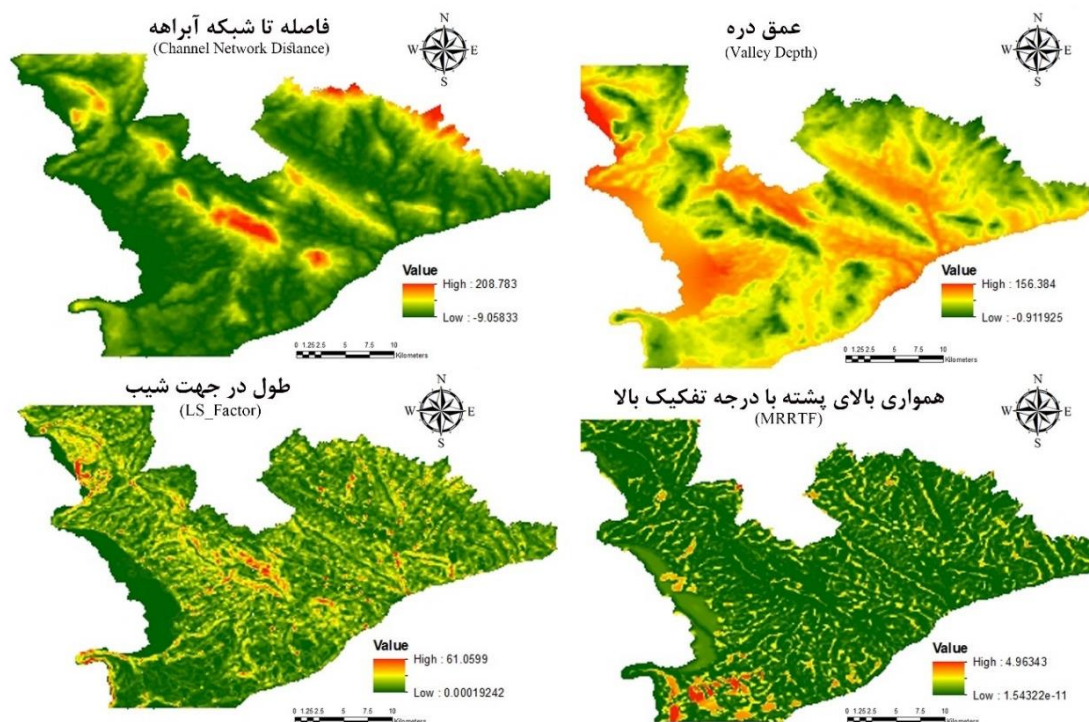
نتایج اعتبارسنجی مدل مورد استفاده در شرایط معمول و بعد از رفع محدودیت داده‌های نامتعادل بر اساس شاخص‌های صحت کلی و شاخص کاپا در جدول ۳ و بر اساس شاخص‌های صحت کاربر و صحت تولیدکننده در جدول ۴ نشان داده شده است. با توجه به نتایج ارائه شده در جدول ۳ در شرایط داده‌های نامتعادل شاخص کاپا برابر ۰/۳۲ و شاخص صحت کلی برابر ۶۵ درصد و پس از متعادل‌سازی داده‌ها با استفاده از دو رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه شاخص کاپا به ترتیب برابر ۰/۵۴ و ۰/۷۷ و شاخص صحت کلی به ترتیب برابر ۷۱ درصد و ۸۶ درصد دقت بالاتری را نسبت به قبل از متعادل‌سازی داده‌ها در پیش‌بینی مکانی کلاس‌های خاک نشان داد. از طرفی بر اساس نتایج به‌دست‌آمده متعادل‌سازی با رویکرد یادگیری حساس به هزینه مقادیر بالاتری را نسبت به رویکرد نمونه‌گیری مجدد برای شاخص‌های صحت‌سنجی کاپا و صحت کلی نشان داد. این امر نشان‌دهنده آن است که در روش یادگیری حساس به هزینه تمرکز مدل بر روی داده‌های با فراوانی کم (اقلیت) است و این امر موجب کاهش خطای

پیش‌بینی و افزایش دقت مدل می‌گردد. مطالعات محققان بسیاری نشان داد بهبود داده‌ها با استفاده از رویکرد نمونه‌گیری مجدد، منجر به افزایش دقت کلی مدل شده و بهبود قابل‌توجهی در حفظ طبقات اقلیت نشان می‌دهند (Sharififar *et al.*, 2019a; Neyestani *et al.*, 2021).

جدایانه از الگوریتم جنگل تصادفی با یادگیری حساس به هزینه استفاده کردند و دریافتند این روش عملکرد و کارایی خوبی در مقایسه با روش‌های معمول جنگل تصادفی دارد و برای مجموعه‌های نمونه کوچک نیز قابل‌استفاده است.



شکل ۳- مقدار اهمیت نسبی متغیرهای کمکی مورداستفاده در پیش‌بینی کلاس‌های خاک در سطح زیرگروه
Fig. 3- The value of the relative importance of covariates used in predicting soil classes at the subgroup level



شکل ۴- نقشه مهم‌ترین متغیرهای محیطی پیش‌بینی کننده کلاس‌های خاک در سطح زیرگروه
Fig. 4- Map of the most important covariates predicting soil classes at the subgroup level

جدول ۱- ویژگی‌های فیزیکی، شیمیایی و طبقه‌بندی خاک‌رخ‌ها در منطقه مورد مطالعه

Table 1. Physical, chemical and classification characteristics of profiles in the study area

افق Horizon	عمق (سانتی‌متر) Depth (cm)	رنگ خاک Soil Color	در صد نسبی ذرات (%)			واکنش خاک pH	قابلیت هدایت الکتریکی EC (dSm-1)	ظرفیت تبادل کاتیونی CEC (Cmol+ kg-1)	کربن آلی OC	کربنات کلسیم معادل Calcium carbonate (%)	گچ Gypsum
			Relative Particle Percentage (%)								
			Top of Form	Bottom of Form							
خاک‌رخ شاهد شماره ۱ تیپیک کلسی زیرپترز Typic Calcixerepts Reference Soil Profile No. 1											
A	0-15	10YR4/4	56	28	16	7.98	0.49	12.90	0.46	16.70	-
Bk1	15-45	10YR5/4	50	22	28	8.30	0.29	18.90	0.26	26.10	-
Bk2	45-65	10YR5/4	64	16	20	8.13	0.29	16.30	0.06	27.90	-
C	65-150	10YR5/4	74	12	14	8.19	0.44	6.40	0.07	25.80	-
خاک‌رخ شاهد شماره ۲ جیپسیک هاپلوزرپترز GypsicHaploxerepts Reference Soil Profile No. 2											
A	0-15	10YR7/2	54	36	10	7.56	2.8	12.50	1.15	21.40	7
Bky1	15-30	10YR8/2	54	36	10	7.86	2.7	9.60	0.61	16.90	23
Bky2	30-47	10YR5/4	46	36	18	7.81	2.9	15.60	0.39	18.90	10
Cky	47-75	7.5YR5/4	52	30	18	7.62	9.26	10.60	0.35	18.30	20
Cy	75-150	7.5YR5/4	58	30	12	7.62	8.10	10.30	0.19	13.10	39
خاک‌رخ شاهد شماره ۳ لیپتیک زراورتنترز Lithic Xerorthents Reference Soil Profile No. 3											
A	0-10	10YR5/6	16	48	36	7.94	0.43	25.60	0.11	19.30	-
C	10-40	10YR5/6	12	48	40	8.77	1.04	23.90	0.13	23.30	-
Cr	40-80	-	16	50	34	8.26	2.53	19.20	0.06	25.50	-
خاک‌رخ شاهد شماره ۴ تیپیک زراورتنترز Typic Xerorthents Reference Soil Profile No. 4											
A	0-20	10YR6/3	68	22	10	7.81	0.60	9.10	0.28	13.10	-
C1	20-45	10YR5/3	68	20	12	7.79	1.01	10.60	0.23	14.30	-
C2	45-80	10YR5/3	68	20	12	7.93	1.03	9.90	0.20	11.50	-
C3	80-150	10YR5/3	80	14	6	8.15	0.65	-	0.10	12.60	-
خاک‌رخ شاهد شماره ۵ تیپیک هاپلوزرپترز Typic Haploxerepts Reference Soil Profile No. 5											
A	0-25	10YR4/4	21	51	28	8.02	0.76	18.00	0.25	21.6	-
Bw1	25-80	10YR4/3	12	42	46	8.63	1.53	28.50	0.16	22.60	-
Bw2	80-150	10YR4/3	8	42	50	8.66	3.77	-	0.17	18.70	-

جدول ۲- تعداد مشاهددها کلاس‌های خاک در سطح زیرگروه
Table 2- The number of observations of soil classes at the subgroup level

کد کلاس Class Code	زیرگروه‌های خاک Subgroups Soil	تعداد مشاهددها Number of Observations	درصد مشاهددها Percentage of Observations
A	تیپیک کلسی‌زریتر Typic Calcixerepts	68	32.43
B	تیپیک هاپلوزریتر Typic Haploxerepts	26	17.56
C	جیپسیک هاپلوزریتر GypsicHaploxerepts	12	8.1
D	تیپیک زراورتنتر Typic Xerorthents	31	20.94
E	لیتیک زراورتنتر Lithic Xerorthents	11	7.34

جدول ۳- صحت پیش‌بینی سطح تاکسونومیک زیرگروه قبل و بعد از متعادل‌سازی داده‌ها (رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه) توسط الگوریتم جنگل تصادفی

Table 3. Prediction accuracy of the taxonomic level of the subgroup before and after data treatment by random forest algorithm (resampling and cost-sensitive learning approaches)

شاخص‌های صحت‌سنجی Validation indicators		رویکردهای متعادل‌سازی داده‌ها Data balancing approaches
ضریب کاپا Kappa coefficient	صحت کلی (%) Overall accuracy (%)	
0.32	65	داده‌های نامتعادل Imbalanced dataset
0.54	71	داده‌های متعادل با رویکرد نمونه‌گیری مجدد Balanced dataset with a resampling approach
0.77	86	داده‌های متعادل با رویکرد یادگیری حساس به هزینه Balanced dataset with a cost-sensitive learning approach

یادگیری حساس به هزینه پیش‌بینی دو کلاس کم رخداد (جیپسیک هاپلوزریتر و لیتیک زراورتنز) با به حداقل رسیدن کل هزینه طبقه‌بندی اشتباه (کم برازش) بهینه‌شده است. نتایج مطالعه Fernández *et al.* (2018) نشان داد که روش‌های یادگیری حساس به هزینه پتانسیل بالایی برای مقابله با مشکل عدم تعادل کلاس‌بندی در داده‌کاوی و یادگیری ماشین دارند. تحقیقات نشان داده است که رویکرد یادگیری حساس به هزینه در برنامه‌هایی که مجموعه داده دارای توزیع کلاس نامتعادل است، عملکرد بهتری را به همراه دارد (Yu *et al.*, 2018). همچنین نتایج محققین دیگری نشان می‌دهد که برای حل مسائل طبقه‌بندی نامتعادل، رویکرد یادگیری حساس به هزینه منجر به

بر اساس نتایج مجموعه داده‌های اعتبارسنجی صحت تولیدکننده و کاربر که در جدول ۴ نشان داده‌شده است، زیرگروه‌های جیپسیک هاپلوزریتر و لیتیک زراورتنز که جزء کلاس‌های اقلیت محسوب می‌شوند هنگام استفاده از کلاس‌های نامتعادل پیش‌بینی نشده و حذف‌شده بودند اما پس از بهبود داده‌ها و بیش‌افزایی با دو رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه به تعداد این دو کلاس اقلیت، پیش‌بینی این زیرگروه‌ها با صحت قابل قبولی افزایش یافت. این در حالی است که بهبود داده‌ها با رویکرد یادگیری حساس به هزینه دقت نسبتاً بالاتری نسبت به رویکرد نمونه‌گیری مجدد در پیش‌بینی مکانی زیرگروه‌های اقلیت خاک نشان داد. این نتایج بیان می‌کند که در رویکرد

عملکرد برتر می‌شود و رویکرد مناسب‌تری نسبت به تکنیک‌های نمونه‌گیری است (Zhang *et al.*, 2019; Mienyeand Sun, 2021).

شکل ۵. محتمل‌ترین نقشه‌های تولیدشده قبل و بعد از متعادل‌سازی داده‌ها با رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه را نشان می‌دهد. با توجه به نتایج صحت تولیدکننده و کاربر (جدول ۴) عملکرد هر دو کلاس اقلیت بهبود قابل‌توجهی را پس از متعادل‌سازی داده‌ها نشان می‌دهد که این امر در نقشه‌های تولیدشده مشهود است. دو کلاس جیپسیک هاپلوزرپتز و لیتیک زراورتنز که هنگام آموزش با استفاده از مجموعه داده نامتعادل حذف‌شده بودند، با صحت کاربر ۱۰۰ و ۷۸ درصد در رویکرد نمونه‌گیری مجدد و ۱۰۰ و ۱۰۰ درصد در رویکرد یادگیری حساس به هزینه با استفاده از داده‌های بهبودیافته، به‌خوبی پیش‌بینی شدند. کلاس‌های اقلیت صحت تولیدکننده ۷۵ و ۸۸ درصد در رویکرد نمونه‌گیری مجدد و ۸۵ و ۹۱ درصد در رویکرد یادگیری حساس به

هزینه را در مقایسه با دقت صفر هنگام آموزش با استفاده از داده‌های نامتعادل نشان دادند. این نتایج بیان می‌کند که ناکافی بودن تعداد مشاهده‌ها در کلاس‌های خاک منجر به پیش‌بینی نادرست و ناموفق الگوریتم برای کلاس‌های اقلیت می‌شود؛ اما پس از افزایش تعداد نمونه‌ها با استفاده از رویکردهای موجود الگوریتم از حذف کلاس‌های اقلیت جلوگیری کرده و در نهایت دقت پیش‌بینی برای کلاس‌های اقلیت بهبود می‌یابد. در واقع تعداد نامتعادل مشاهده‌های کلاس‌های خاک می‌تواند تأثیر منفی بر نتایج اعتبارسنجی داشته باشد (Rahimi *et al.*, 2023b). نتایج اعتبارسنجی مدل‌های مختلف در مطالعات مختلف نقشه‌برداری رقومی نشان می‌دهد که صرف‌نظر از نوع رویکرد به کار گرفته‌شده، نقشه‌های تهیه‌شده با استفاده از بهبود داده‌های نامتعادل دارای دقت بالاتری نسبت به نقشه‌های تهیه‌شده با داده‌های معمول بوده‌اند (Taghizadeh-Mehrjardi *et al.*, 2020; Sharififar *et al.*, 2019b).

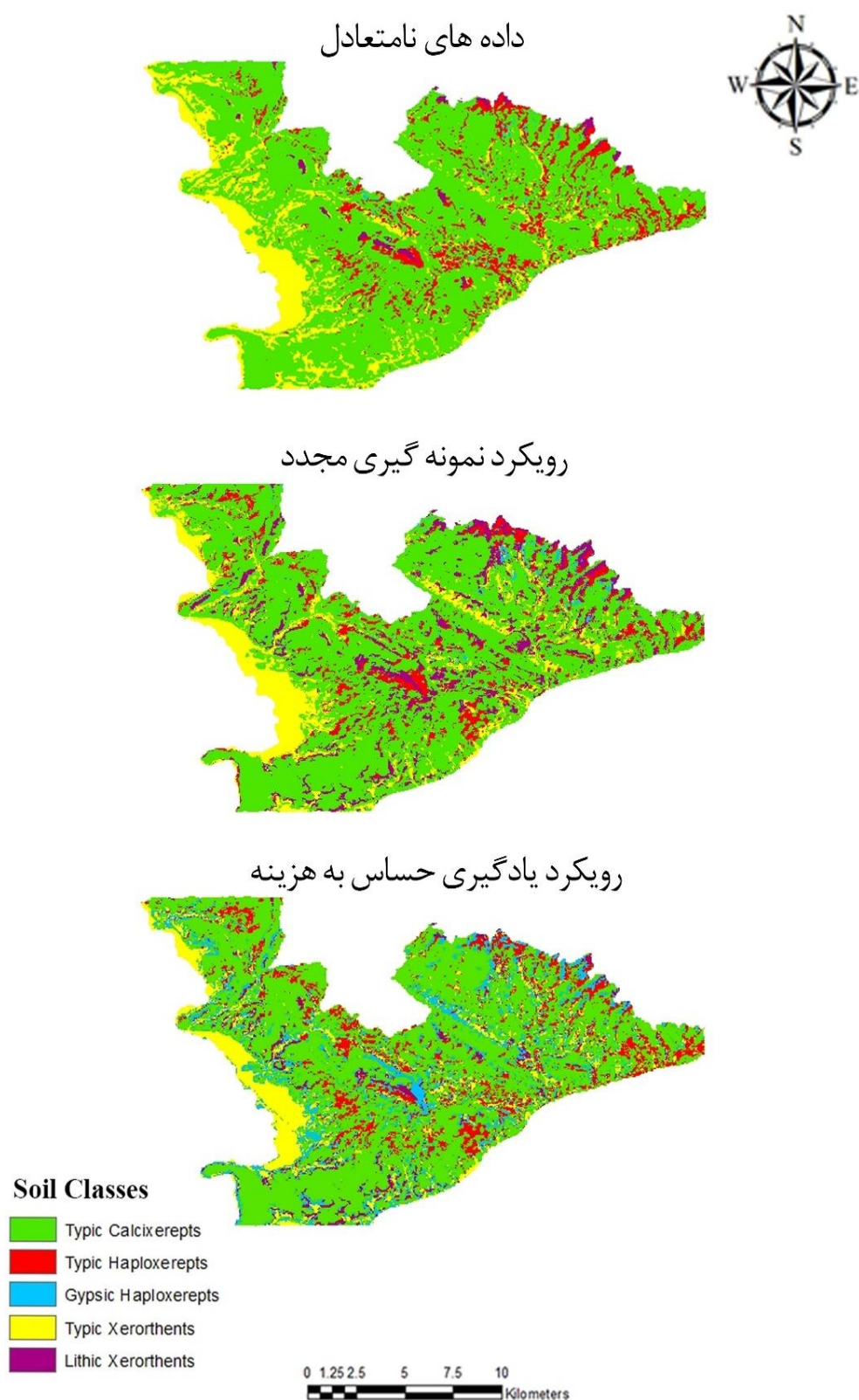
جدول ۴- صحت تولیدکننده و کاربر برای کلاس‌های خاک در سطح زیرگروه قبل و بعد از متعادل‌سازی داده‌ها (رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه) بر اساس مدل جنگل تصادفی

Table 4- Producer and User accuracy of soil classes at the subgroup level before and after data treatment based on the random forest (resampling and cost-sensitive learning approaches)

صحت کاربر (%) User accuracy (%)		صحت تولیدکننده (%) Producer accuracy (%)		قابلیت اطمینان Validation	
داده‌های متعادل Balanced dataset		داده‌های نامتعادل Imbalanced dataset		مدل‌های یادگیری ماشین Machine learning models	
نمونه‌گیری مجدد Resampling	یادگیری حساس به هزینه Cost-sensitive learning	نمونه‌گیری مجدد Resampling	یادگیری حساس به هزینه Cost-sensitive learning	کلاس‌های خاک در سطح زیرگروه Subgroup of soil	
75	95	61	85	85	A
34	71	100	25	83	50
100	100	NaN	75	85	0
34	81	65	34	100	34
78	100	NaN	88	91	0

NaN*: عدد نیست، هیچ پیش‌بینی برای این کلاس انجام نشده است.

NaN*: Not a number, no prediction has been made for this class.



شکل ۵- نقشه‌های تولیدشده توسط الگوریتم جنگل تصادفی قبل و بعد از متعادل‌سازی داده‌ها با دو رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه

Fig. 5- Maps produced by random forest algorithm before and after data balancing with resampling and cost-sensitive learning approaches

نتیجه گیری

این پژوهش باهدف کارآمدی روش‌های پیش‌درمانی یا بهبود داده‌های نامتعادل برای تولید نقشه‌های خاک با دقت بالاتر و مدیریت بهتر اراضی کشاورزی انجام گرفت. در این مطالعه دو رویکرد یادگیری حساس به هزینه و نمونه‌گیری مجدد برای حل مسئله عدم تعادل داده‌ها مورد استفاده قرار گرفت. نتایج نشان داد که روش‌های بهبود داده‌های نامتعادل می‌تواند سبب افزایش دقت مدل در تشخیص کلاس‌های اقلیت شده و در نتیجه منجر به افزایش دقت پیش‌بینی مکانی کلاس‌های خاک و تهیه نقشه‌های دقیق‌تر گردد. از طرفی رویکرد یادگیری حساس به هزینه با تمرکز بر روی کلاس‌های با تکرار کم، نشان داد که می‌تواند به‌عنوان یک مدل برتر در دیگر مناطق نیز مورد استفاده قرار گیرد. با توجه به اینکه تحقیقات در زمینه داده‌های نامتعادل در خاک محدود است، این مطالعه می‌تواند یک راه‌حل مؤثر برای مقابله با داده‌های نامتعادل در کلاس‌های خاک و تولید نقشه‌های رقومی خاک با دقت بالا باشد.

سپاسگزاری

نویسندگان مراتب سپاس و قدردانی خود را از حمایت‌های معنوی مؤسسه تحقیقات خاک و آب کشور ابراز می‌دارند.

پی‌نوشت‌ها

- ¹ Resampling Method
- ² Ensemble Models
- ³ Hill lands
- ⁴ Piedmont plains
- ⁵ Mesic
- ⁶ Xeric
- ⁷ Zinck
- ⁸ Principal Component Analysis, PCA
- ⁹ Breiman
- ¹⁰ Overall Accuracy, OA
- ¹¹ Producer Accuracy, PA
- ¹² Users Accuracy, UA
- ¹³ Kappa Index
- ¹⁴ Analytical HillShading
- ¹⁵ Sunrise
- ¹⁶ Valley Depth
- ¹⁷ LS_Factor
- ¹⁸ Channel Network Distance
- ¹⁹ Topographic Wetness Index, TWI
- ²⁰ Multi-Resolution Ridge Top Flatness Index, MRRTF
- ²¹ Mean Decrease Accuracy

References

- Abdi, L., & Hashemi, S. (2015). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE transactions on Knowledge and Data Engineering*, 28(1), 238-251. DOI: 10.1109/TKDE.2015.2458858.
- Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781315139470>.
- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
- Breiman, L., & Cutler, A. (2004). *Random Forests*. Department of Statistics, University of Berkeley.
- Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards Jr, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, 239, 68-83. <https://doi.org/10.1016/j.geoderma.2014.09.019>.
- Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1), 35-46. [doi.org/10.1016/00344257\(91\)90048-B](https://doi.org/10.1016/00344257(91)90048-B).

منابع

- Devi, D., Biswas, S. K., & Purkayastha, B. (2019). A cost-sensitive weighted random forest technique for credit card fraud detection. In 2019 10th international conference on computing, communication and networking technologies (ICCCNT). 1-6. doi:10.1109/ICCCNT45670.2019.8944885.
- Fantappiè, M., L'Abate, G., Schillaci, C., & Costantini, E. A. (2023). Digital soil mapping of Italy to map derived soil profiles with neural networks. *Geoderma Regional*, 32, e00619. <https://doi.org/10.1016/j.geodrs.2023.e00619>.
- Fernández, A., del Jesus, M. J., & Herrera, F. (2009). On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets. *Expert Systems with Applications*, 36(6), 9805-9812. doi.org/10.1016/j.eswa.2009.02.048.
- Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, 265, 62-77. <https://doi.org/10.1016/j.geoderma.2015.11.014>.

- Jensen, J. R. (2005). Introductory image processing: A remote sensing perspective.
- Kang, M., Liu, Y., Wang, M., Li, L., & Weng, M. (2022). A random forest classifier with cost-sensitive learning to extract urban landmarks from an imbalanced dataset. *International Journal of Geographical Information Science*, 36(3), 496-513. doi.org/10.1080/13658816.2021.1977814.
- Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, 81, 401-418. https://doi.org/10.1016/j.apm.2019.12.016.
- Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in artificial intelligence*, 5(4), 221-232.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.
- Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 154(1-2), 138-152. https://doi.org/10.1016/j.geoderma.2009.10.007.
- McBratney, A.B., Santos, M.M. and Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1-2):3-52. https://doi.org/10.1016/S0016-7061(03)00223-4.
- Meng, X. T., Yan, F. G., Cao, B. X., Jin, M., & Zhang, Y. (2022). Efficient real-valued DOA estimation based on the trigonometry multiple angles transformation in monostatic MIMO radar. *Digital Signal Processing*, 123, 103437. Mienye, I.D. and Sun, Y., 2021. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25:100690. https://doi.org/10.1016/j.dsp.2022.103437.
- Minasny, B., & Hartemink, A. E. (2011). Predicting soil properties in the tropics. *Earth-Science Reviews*, 106(1-2), 52-62. https://doi.org/10.1016/j.earscirev.2011.01.005.
- Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in agronomy*, 118, 1-47. https://doi.org/10.1016/B978-0-12-405942-9.00001-3.
- Mousavi, S. R., Sarmadian, F., & Rahmani, A. (2020). Modelling and prediction of soil classes using boosting regression tree and random forests machine learning algorithms in some part of Qazvin plain. *Iranian Journal of Soil and Water Research*, 50(10), 2525-2538. https://dx.doi.org/10.22059/ijswr.2019.280905.668198.
- Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A., & Sharififar, A. (2021). Digital mapping of soil classes using spatial extrapolation with imbalanced data. *Geoderma Regional*, 26, e00422. https://doi.org/10.1016/j.geodrs.2021.e00422.
- Rahimi mashkale, M., Delavar, M. A., jamshidi, M., & sharififar, A. (2023a). Improving the classification of Soil imbalanced data using machine learning algorithms in Some Part of Zanjan province land. *Journal of Agricultural Engineering Soil Science and Agricultural Mechanization, Scientific Journal of Agriculture*, 46(1), 61-82. doi: 10.22055/AGEN.2023.43838.1667.[In Persian]
- Rahimi Mashkaleh, M., amirdelavar, M., jamshidi, M., & sharififar, A. (2023b). Modeling Spatial Distribution of Soil Classes Using Machine Learning Algorithms in Some Parts of Zanjan Province. *Iranian Journal of Soil Research*, 37(2), 147-165. doi: 10.22092/ijsr.2023.361649.698. [In Persian]
- Schoeneberger, P. J., Wysocki, D. A., & Benham, E. C. (Eds.). (2012). *Field book for describing and sampling soils*. Government Printing Office.
- Sharififar, A., Sarmadian, F., & Minasny, B. (2019a). Mapping imbalanced soil classes using Markov chain random fields models treated with data resampling technique. *Computers and Electronics in Agriculture*, 159, 110-118. doi.org/10.1016/j.compag.2019.03.006.
- Sharififar, A., Sarmadian, F., Malone, B. P., & Minasny, B. (2019b). Addressing the issue of digital mapping of soil classes with imbalanced class observations. *Geoderma*, 350, 84-92. doi.org/10.1016/j.geoderma.2019.05.016.
- Sharififar, A., & Sarmadian, F. (2023). Coping with imbalanced data problem in digital mapping of soil classes. *European Journal of Soil Science*, 74(3), e13368. doi.org/10.1111/ejss.13368.
- Soil and Water Research Institute. (2010). *Site Selection, Soil Survey and Land Evaluation for Development of Orchards in Zanjan Province, Iran*. [In Persian]
- Soil science division staff. (2017). *Soil survey manual*". USDA Handbook 18: 120-131.
- Soil Survey Staff. (2022). *Keys to soil taxonomy*, 13th edition. USDA Natural Resources Conservation Service.
- Statistical Yearbook of Zanjan Province. (2019). *Land and Climate*, National Statistics Organization. [In Persian]
- Taghizadeh-Mehrjardi, R., Minasny, B., Toomanian, N., Zeraatpisheh, M., Amirian-Chakan, A., & Triantafilis, J. (2019). Digital mapping of soil classes using ensemble of models in Isfahan region, Iran. *Soil Systems*, 3(2), 37. https://doi.org/10.3390/soilsystems3020037.
- Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian,

N., Scholten, T., & Schmidt, K. (2020). Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. *Geoderma*, 376, 114552. doi.org/10.1016/j.geoderma.2020.114552.

Usda, N. R. C. S. (2004). Soil survey laboratory methods manual. Soil survey investigations report, 42.

Vincent, S., Lemerrier, B., Berthier, L., & Walter, C. (2018). Spatial disaggregation of complex Soil Map Units at the regional scale based on soil landscape relationships. *Geoderma*, 311, 130-142. doi.org/10.1016/j.geoderma.2016.06.006.

Wadoux, A. M. C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210, 103359. doi.org/10.1016/j.earscirev.2020.103359.

Yu, H., Sun, C., Yang, X., Zheng, S., Wang, Q., & Xi, X. (2018). LW-ELM: a fast and flexible cost-sensitive learning framework for classifying

imbalanced data. *IEEE Access*, 6, 28488-28500. doi: 10.1109/ACCESS.2018.2839340.

Zhang, C., Tan, K. C., Li, H., & Hong, G. S. (2018). A cost-sensitive deep belief network for imbalanced classification. *IEEE transactions on neural networks and learning systems*, 30(1), 109-122. DOI: 10.1109/TNNLS.2018.2832648.

Zhang, C., Wang, X., Liu, J., Li, M., & Zhang, J. (2021). Cost-sensitive soil mapping using a deep learning approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179, 172-183.

Zinck, J. A., Metternicht, G., Bocco, G., & Del Valle, H. (2016). *Geopedology. An integration of geomorphology and pedology for soils and landscape studies*: Springer International Publishing Switzerland, 556p.



*This page is intentionally
left blank.*