Original Article

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

Mastaneh Rahimi Mashkaleh,[1] Mohammad Amir Delavar,[1*] Mohammad Jamshidi[2]

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

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

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, Gypsic Haploxerepts, 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. Gypsic Haploxerepts and Lithic Xerorthents subgroups, considered minority classes, were unidentified and excluded when using imbalanced classes.

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

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.

Keywords: Random forest, Imbalanced data, Resampling, Cost-sensitive learning

مقاله پژوهشی

کارایی رویکردهای متعادل‌سازی داده در نقشه‌برداری رقومی خاک (مطالعه موردی: بخشی از اراضی استان زنجان)

مستانه رحیمی مشکله[1]، محمد امیر دلاور[1*]، محمد جمشیدی[2]

[1] گروه علوم خاک، دانشکده کشاورزی، دانشگاه زنجان، زنجان، ایران

[2] موسسه تحقیقات خاک و آب، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران

سابقه و هدف: نقشه‌برداری رقومی خاک با استفاده از روش‌های نوین یادگیری ماشین به‌طور گسترده‌ای برای پیش‌بینی پراکندگی مکانی و ویژگی‌های مختلف خاک به کار گرفته می‌شود، بااین‌وجود یکی از محدودیت‌های استفاده از روش‌های نقشه‌برداری رقومی خاک در مطالعات خاکشناسی، عدم تعادل کلاس‌های خاک است که تأثیر منفی بر عملکرد الگوریتم‌های یادگیری ماشین دارد؛ بنابراین این پژوهش برای رفع این چالش و بهبود طبقه‌بندی کلاس‌های نامتعادل خاک با دو رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه و استفاده از مدل پیش‌بینی جنگل تصادفی در استان زنجان انجام‌گرفته است.

مواد و روش‌ها: تعداد ۱۴۸ خاک‌رخ مشاهداتی بر اساس الگوی طبقه‌بندی تصادفی با فاصله ۵۰۰ متر حفر و پس از انتقال به آزمایشگاه تجزیه‌های مختلف فیزیکی و شیمیایی مطابق با روش‌های استاندارد بر روی آن‌ها انجام گرفت. متغیرهای محیطی شامل اطلاعات نقشه‌های ژئومورفولوژی و زمین‌شناسی، مدل رقومی ارتفاع و داده‌های حاصل از تصاویر ماهواره‌ای لندست ۸ بودند که بر اساس نظر کارشناسی و رویکرد تحلیل مؤلفه اصلی تعدادی از متغیرهای محیطی شامل اطلاعات نقشه‌های ژئومورفولوژی، اطلاعات زمین‌شناسی و ویژگی‌های مستخرج از مدل رقومی ارتفاع به‌عنوان مؤثرترین متغیرهای محیطی برای پیش‌بینی کلاس‌های خاک و به‌عنوان ورودی مدل انتخاب شدند. مدل‌سازی رابطه خاک - زمین‌نما با استفاده از الگوریتم جنگل تصادفی و اصلاح داده‌های نامتعادل توسط رویکرد نمونه‌گیری 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 برنامه‌نویسی R_{studio} انجام شد.

نتایج و بحث: نتایج این پژوهش حاکی از این بود که خاک‌های منطقه در سطح زیرگروه در پنج کلاس با توزیع نامتعادل شامل تیپیک کلسی‌زرپتز، تیپیک هاپلوزرپتز، جیپسیک هاپلوزرپتز، تیپیک زراورتنتز و لیتیک زراورتنتز طبقه‌بندی شدند. بر این اساس مقادیر آماره‌های صحت

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

کلی و ضریب کاپا برای ارزیابی نقشه خاک با داده‌های نامتعادل به ترتیب برابر ۶۵ درصد و ۰/۳۲ بوده و پس از متعادل‌سازی داده‌ها در رویکرد نمونه‌گیری 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. زیرگروه‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنز که جزء کلاس‌های اقلیت محسوب می‌شدند، هنگام استفاده از کلاس‌های نامتعادل پیش‌بینی‌نشده 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 نشان داد.

نتیجه‌گیری: نتایج ارزیابی مدل‌ها نشان داد که در مدل‌سازی با استفاده از توزیع نامتعادل کلاس‌های خاک، به دلیل از دست رفتن کلاس‌های با تعداد مشاهده کم، نقشه‌های نامطمئن با دقت نسبتاً ضعیفی ایجاد می‌شود که پس از اعمال متعادل‌سازی داده‌ها، دقت مدل‌های مبتنی بر روابط خاک - زمین‌نما در مطالعات نقشه‌برداری رقومی خاک ارتقا می‌یابد. نتایج نشان داد که رویکرد یادگیری حساس به هزینه با تمرکز بر روی کلاس‌های با تکرار کم، می‌تواند به‌عنوان یک مدل برتر در مناطق دیگر نیز مورداستفاده قرار گیرد. با توجه به اینکه تحقیقات درزمینه 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.

واژه‌های کلیدی: جنگل تصادفی، داده‌های نامتعادل، نمونه‌گیری مجدد، یادگیری حساس به هزینه

مقدمه

نقشه‌برداری رقومی خاک با استفاده از فن‌های نوآورانه یادگیری ماشین به‌طور فزاینده‌ای در پیش‌بینی توزیع مکانی و ویژگی‌های مختلف خاک به کار می‌رود (Khaledian and Miller, 2020; Heung *et al*., 2016; Brungard *et al*., 2015). 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 توسعه‌یافته‌اند تا اطلاعات و استنباط‌های خاک را به‌صورت پیوسته ارائه دهند (Malone *et al*., 2009; Minasny *et al*., 2013; McBratney *et al*., 2003). 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 (Wadoux *et al*., 2020; Minasny and Hartemink, 2011).

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 (Taghizadeh-Mehrjardi *et al*., 2020; Heung *et al*., 2016).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. به‌عبارت‌دیگر، این پدیده ممکن است باعث حذف یک ناحیه از منطقه موردمطالعه با تعداد مشاهده‌های کم‌تر شود. درواقع 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 ارائه‌شده است می‌توان نمونه‌گیری مجدد* و یادگیری حساس به هزینه† را نام برد (Rahimi et al., 2023a; Sharififar et al., 2019a; Taghizadeh-Mehrjardi et al., 2019). 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 (Sharififar et al., 2019b; Vincent et al., 2018).

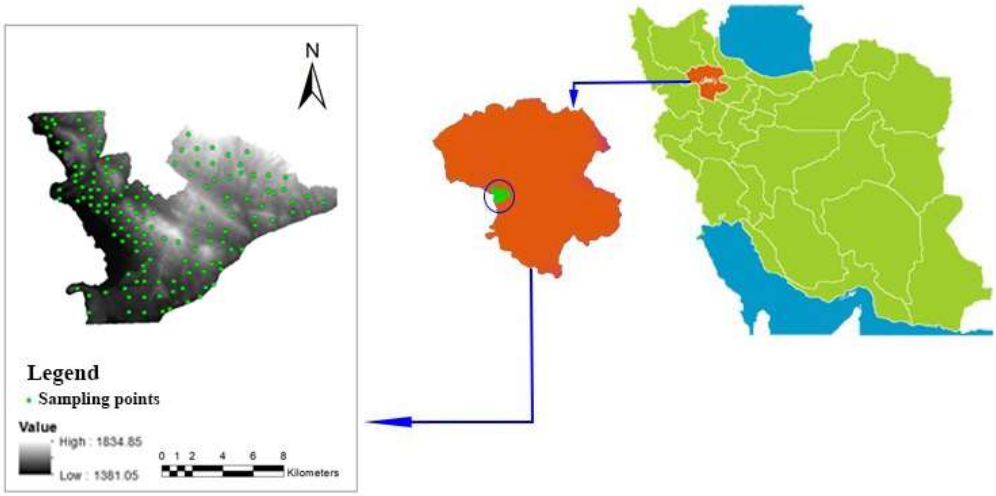
باوجوداینکه نقشه‌برداری رقومی خاک به‌عنوان یک استراتژی مؤثر برای افزایش بهره‌وری خاک در تولید پایدار مورداستفاده قرار می‌گیرد. اما نامتوازن بودن داده‌ها در این مطالعات به‌عنوان یک چالش در نظر گرفته می‌شود. بنابراین این پژوهش باهدف دستیابی به تحلیل دقیق‌تر و کاربردی‌تر داده‌های خاک و تولید نقشه‌های با 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.

*. Resampling Method

†. Ensemble Models

مواد و روش‌ها

منطقه موردمطالعه 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 واقع‌شده است (شکل ۱). میانگین ارتفاع منطقهٔ موردمطالعه از سطح دریا ۱۴۸۲ متر است. با توجه به آمار بلندمدت ۲۰ ساله، متوسط بارندگی سالانه منطقه ۳۴۰ میلی‌متر 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 تپه‌ماهور* و دشت‌های دامنه‌ای† و عمده سازندهای زمین‌شناسی منطقه شامل لایه‌های کربناته، سنگ‌آهک، کنگلومرا و مواد آتش‌فشانی و آذرآواری مانند توف اسیدی است. منطقه دارای پوشش گیاهی تنک است و در زمره مراتع ضعیف قرارمی گیرد. با استناد به نقشه رژیم‌های رطوبتی و حرارتی ایران، منطقه مذکور دارای رژیم حرارتی مزیک‡ و رژیم رطوبتی زریک§ هستند (Soil and Water Research Institute, 2010).



شکل ۱- موقعیت منطقه موردمطالعه و نقاط نمونه‌برداری
Figure 1- Location of the study area and sampling points

*. Hill lands

†. Piedmont plains

‡. Mesic

§. Xeric

عملیات صحرایی و تجزیه‌های آزمایشگاهی

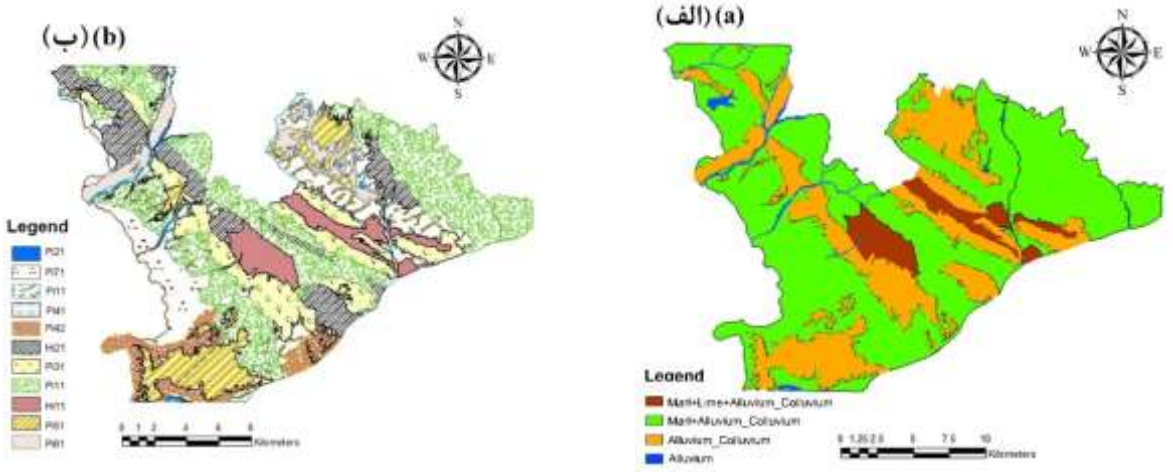
مطابق روش‌های استاندارد مطالعات خاک‌شناسی تعداد ۱۴۸ خاک‌رخ بر اساس الگوی طبقه‌بندی تصادفی با میانگین فاصله ۵۰۰ متر (در برخی نقاط بر اساس نظر کارشناس و شرایط محلی تا فاصله ۷۰۰ متر تغییر یافتند)، حفر شد (Soil science division staff, 2017). نمونه‌برداری و تشریح تمامی خاک‌رخ‌های حفرشده بر اساس راهنمای تشریح خاک‌رخ‌ها انجام گردید (Schoeneberger et al., 2012). نمونه‌ها پس از هوا خشک شدن از الک ۲ میلی‌متری عبور داده شدند و تجزیه‌های فیزیکی و شیمیایی با استفاده از روش‌های استاندارد و متداول انجام گرفت (USDA, 2004). خاک‌رخ‌ها بر اساس سیستم جامع رده‌بندی خاک به روش آمریکایی تا سطح فامیل طبقه‌بندی شدند (Soil Survey Staff, 2022).

استخراج متغیرهای محیطی

از اطلاعات نقشه‌های ژئومورفولوژی، نقشه زمین‌شناسی، داده‌های سنجش‌ازدور و مدل رقومی ارتفاع باقدرت تفکیک مکانی ۳۰ متر (سنجنده استر) برای استخراج لایه‌های کمکی خاک استفاده شد. نقشه ژئومورفولوژی منطقه بر اساس تلفیق لایه‌های اطلاعاتی شامل واحدهای لندفرم و مواد مادری به همراه تفسیر تصاویر ماهواره‌ای بر اساس رویکرد سلسله مراتبی ارائه‌شده توسط زینک* (۲۰۱۵) تهیه گردید (جدول ۱، شکل ۲_الف). نقشه زمین‌شناسی با مقیاس ۱:۲۵۰۰۰۰ منطقه تهیه‌شده توسط سازمان زمین‌شناسی کشور در محیط Arc-GIS (نسخه ۱۰/۷) زمین مرجع و رقومی شد (شکل ۲_ب). ۳۶ متغیر سنجش‌ازدوری با استفاده از تصاویر سنجنده (OLI/TIRS) ماهواره لندست ۸ باقدرت تفکیک مکانی ۳۰×۳۰ متر (USGS 2014) پس از اعمال تصحیحات رادیومتریکی و اتمسفری در محیط نرم‌افزار ENVI (نسخه ۵/۳) و ۱۸ متغیر پستی‌وبلندی از مدل رقومی ارتفاع در محیط نرم‌افزار SAGA GIS (نسخه ۷/۹) استخراج شد. انتخاب متغیرهای کمکی در این مطالعه بر اساس رویکرد تحلیل مؤلفه اصلی† در نرم‌افزار SPSS (نسخه ۲۶) و رتبه‌بندی اهمیت نسبی مدل یادگیری ماشین به همراه نظارت کارشناس انجام شد (Kuhn and Johnson, 2013).

*. Zinck

†. Principal Component Analysis, PCA



شکل ۲- الف) نقشه زمین‌شناسی ب) نقشه ژئومورفولوژی منطقه موردمطالعه

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

جدول ۱- واحدهای تفکیک‌شده در سطح لندفرم بر اساس اطلاعات ژئومورفولوژی در منطقه موردمطالعه

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**	سنگ‌شناسی /منشأ **Lithology/Origin**	پستی‌وبلندی /قالب **Relief/Morphology**	زمین‌نما **Physiography**
Consociation	Hi111	Slope facet complex	Marl + Lime + Alluvium - Colluvium	Medium hill	تپه‌ماهور Hill lands
Complex	Hi211 Hi212 Hi213 Hi214 Hi215	Summit Shoulder Backslope Footslope Toeslope	Marl + Alluvium - Colluvium	Low hill	
Association	Pi111	High glacic	Alluvium - Colluvium	Glacic, Dissected	دشت دامنه‌ای **Piedmont plains**
Complex	Pi211	Middle glacic	Marl + Alluvium - Colluvium	Glacic, Moderately dissected	
Association	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 Pi712 Pi713	Upper part Middle part Lower part	Marl + Alluvium - Colluvium	Channeled recent alluvial deposits	

مدل‌سازی خاک_ زمین‌نما

مدل جنگل تصادفی: مدل جنگل تصادفی یک روش ناپارامتری است که اولین بار توسط بریمن* و همکاران (۱۹۸۴) ارائه شد. این مدل قادر به پیش‌بینی متغیرهای کمّی یا متغیرهای طبقه‌بندی‌شده بر اساس مجموعه‌ای از متغیرهای پیش‌بینی کننده کمّی 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 R_{Studio} انجام شد.

متعادل‌سازی داده با استفاده از رویکرد نمونه‌گیری مجدد از داده‌ها: در این رویکرد برای اصلاح داده‌های نامتعادل در دو مرحله بیش‌نمونه‌گیری 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 نکند. این روش‌ها باعث می‌شوند که توزیع داده‌های کلاس‌های خاک متعادل شود (Abdi and Hashemi, 2015).

متعادل‌سازی داده‌با استفاده از رویکرد یادگیری حساس‌به 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 مورداستفاده باعث می‌شود که توزیع کلاس‌ها در داده‌های آموزشی موردتوجه 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 کم رخداد (اقلیت) وزن بالاتری در مدل داشته باشند و تأثیر آن‌ها در پیش‌بینی‌های مدل بیشتر شود. در این مطالعه، از

*. Breiman

تابع rf در بسته Random Forest در محیط برنامه‌نویسی R_{studio} برای پیاده‌سازی رویکرد حساس به هزینه استفاده شد.

اعتبارسنجی دقت مدل

به‌منظور ارزیابی دقت مدل مورداستفاده، 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*، دقت تولیدکننده†، دقت کاربر‡ و ضریب کاپا§ برای اعتبارسنجی به‌کار گرفته شدند (Jensen, 1996; Congalton, 1991).

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 تجزیه‌وتحلیل سایه‌اندازی تپه‌ها**، طلوع خورشید††، عمق دره‌ ‡‡، شاخص طول در جهت شیب§§، فاصله تا شبکه آبراهه***، شاخص رطوبتی توپوگرافی†††و شاخص همواری بالای پشته با درجه تفکیک بالا‡‡‡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

*. Overall Accuracy, OA

†. Producer Accuracy, PA

‡. Users Accuracy, UA

§. Kappa Index

**. Analytical Hill Shading

††. Sunrise

‡‡. Valley Depth

§§. LS_Factor

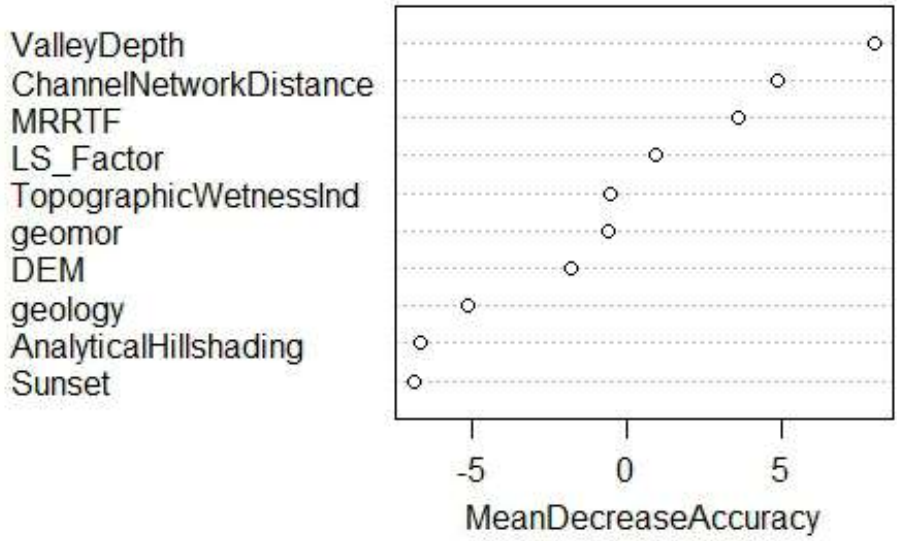
***. Channel Network Distance

†††. Topographic Wetness Index, TWI

‡‡‡. Multi-Resolution Ridge Top Flatness Index, MRRTF

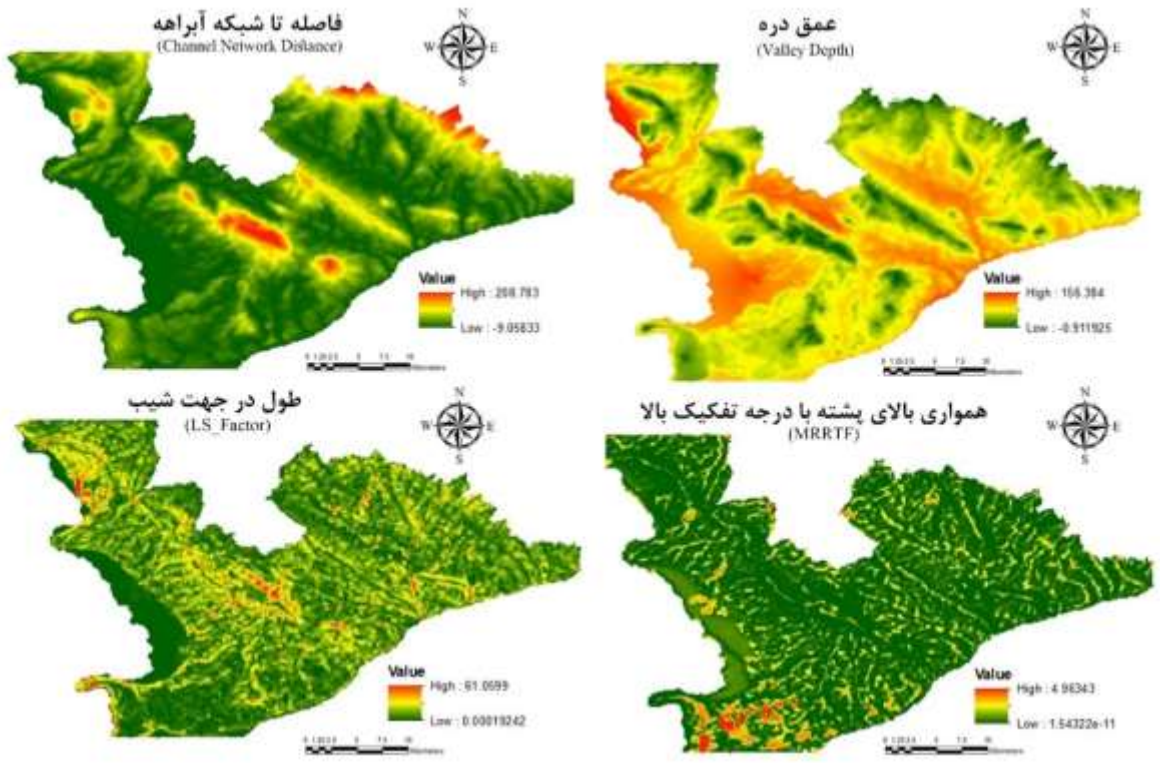
§§§. Mean Decrease Accuracy

در مدل‌سازی مکانی خاک‌ها در منطقه مطالعاتی بوده است (شکل ۳). پراکنش مکانی چهار مورد از مهم‌ترین متغیرهای محیطی در پیش‌بینی‌کننده کلاس‌های خاک در سطح زیرگروه توسط مدل جنگل تصادفی در شکل ۴ نشان داده‌شده 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 موردمطالعه است. Mousavi et al. (2020) در مطالعه خود گزارش کردند که پارامترهای توپوگرافی به‌عنوان مهم‌ترین پیشران‌های محیطی برای مدل‌سازی کلاس‌های خاک هستند.



شکل ۳- مقدار اهمیت نسبی متغیرهای کمکی مورداستفاده در پیش‌بینی کلاس‌های خاک در سطح زیرگروه

Figure 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
Figure 4- Map of the most important covariates predicting soil classes at the subgroup level

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 (Gypsic Haploxerepts)، تیپیک زراورتنتز (Typic Xerorthents) و لیتیک زراورتنتز (Lithic Xerorthents) طبقه‌بندی شدند (Rahimi et al., 2023a). نتایج پیش‌بینی مکانی کلاس‌های خاک منطقه در جدول ۲ نشان داده‌شده است. نتایج این جدول نشان می‌دهد که زیرگروه‌های جیپسیک هاپلوزرپتز (کلاس C) و لیتیک زراورتنتز (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 زراورتنتز (کلاس D) به ترتیب با فراوانی بیشتر از ۳۲/۴۳، ۱۷/۵۶ و ۲۰/۹۴ درصد از کل مشاهده‌های منطقه به‌عنوان کلاس‌های پُرتعداد (اکثریت) هستند.

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

Table 1- Physical, chemical and classification characteristics of profils in the study area

گچ Gypsum	آهک Calcium carbonate	کربن آلی OC	ظرفیت تبادل کاتیونی CEC (Cmol+ kg^{-1})	قابلیت هدایت الکتریکی EC (dSm^{-1})	واکنش خاک pH	در صد نسبی ذرات (٪) Relative Particle Percentage (%)Top of FormBottom of Form			رنگ خاک Soil Color	عمق (سانتی‌متر) Depth (cm)	افق Horizon
	(%)					رس Clay	سیلت Silt	شن Sand			
خاک‌رخ شاهد شماره ۱ تیپیک کلسی زرپتز Reference Soil Profile No. 1 Typic Calcixerepts											
-	16.70	0.46	12.90	0.49	7.98	16	28	56	10YR4/4	0-15	**A**
-	26.10	0.26	18.90	0.29	8.30	28	22	50	10YR5/4	15-45	**Bk1**
-	27.90	0.06	16.30	0.29	8.13	20	16	64	10YR5/4	45-65	**Bk2**
-	25.80	0.07	6.40	0.44	8.19	14	12	74	10YR5/4	65-150	**C**
خاک‌رخ شاهد شماره ۲ جیپسیک هاپلوزرپتز Reference Soil Profile No. 2 Gypsic Haploxerepts											
7	21.40	1.15	12.50	2.8	7.56	10	36	54	10YR7/2	0-15	**A**
23	16.90	0.61	9.60	2.7	7.86	10	36	54	10YR8/2	15-30	**Bky1**
10	18.90	0.39	15.60	2.9	7.81	18	36	46	10YR5/4	30-47	**Bky2**
20	18.30	0.35	10.60	9.26	7.62	18	30	52	7.5YR5/4	47-75	**Cky**
39	13.10	0.19	10.30	8.10	7.62	12	30	58	7.5YR5/4	75-150	**Cy**
خاک‌رخ شاهد شماره ۳ لیتیک زراورتنتز Reference Soil Profile No. 3 Lithic Xerorthents											
-	19.30	0.11	25.60	0.43	7.94	36	48	16	10YR5/6	0-10	**A**
-	23.30	0.13	23.90	1.04	8.77	40	48	12	10YR5/6	10-40	**C**
-	25.50	0.06	19.20	2.53	8.26	34	50	16	-	40-80	**Cr**
خاک‌رخ شاهد شماره ۴ تیپیک زراورتنتز Reference Soil Profile No. 4 Typic Xerorthents											
-	13.10	0.28	9.10	0.60	7.81	10	22	68	10YR6/3	0-20	**A**
-	14.30	0.23	10.60	1.01	7.79	12	20	68	10YR5/3	20-45	**C1**
-	11.50	0.20	9.90	1.03	7.93	12	20	68	10YR5/3	45-80	**C2**

-	12.60	0.10	-	0.65	8.15	6	14	80	10YR5/3	80-150	**C3**
Reference Soil Profile No. 5 Typic Haploxerepts خاک‌رخ شاهد شماره ۵ تیپیک هاپلوزرپتز											
-	21.6	0.25	18.00	0.76	8.02	28	51	21	10YR4/4	0-25	**A**
-	22.60	0.16	28.50	1.53	8.63	46	42	12	10YR4/3	25-80	**Bw1**
-	18.70	0.17	-	3.77	8.66	50	42	8	10YR4/3	80-150	**Bw2**

جدول ۲- تعداد مشاهده‌ها 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 مشاهده‌ها Number of Observations	درصد مشاهده‌ها Percentage of Observations
A	تیپیک کلسی‌زرپتز Typic Calcixerepts	68	32.43
B	تیپیک هاپلوزرپتز Typic Haploxerepts	26	17.56
C	جیپسیک هاپلوزرپتز Gypsic Haploxerepts	12	8.1
D	تیپیک زراورتنتز Typic Xerorthents	31	20.94
E	لیتیک زراورتنتز Lithic Xerorthents	11	7.34

نتایج اعتبارسنجی مدل مورداستفاده 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 داده‌شده است. با توجه به نتایج ارائه‌شده 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 (Neyestani et al., 2021; Sharififar et al., 2019a). Kang et al. (2022) و Devi et al. (2019) نیز در مطالعاتی جداگانه 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.

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 datase with a resampling approach
0.77	86	داده‌های متعادل با رویکرد یادگیری حساس به هزینه Balanced dataset with a cost-sensitive learning approach

بر اساس نتایج مجموعه داده‌های اعتبارسنجی صحت تولیدکننده و کاربر که در جدول ۴ نشان داده‌شده است، زیرگروه‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنز که جزء کلاس‌های اقلیت محسوب می‌شوند هنگام استفاده از کلاس‌های نامتعادل پیش‌بینی‌نشده 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 خاک‌نشان داد. این نتایج بیان می‌کند که در رویکرد یادگیری حساس به هزینه پیش‌بینی دو کلاس کم رخداد (جیپسیک هاپلوزرپتز 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 عملکرد برتر می‌شود و رویکرد مناسب‌تری نسبت به تکنیک‌های نمونه‌گیری است (Mienye and Sun, 2021; Zhang et al., 2019).

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. درواقع 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).

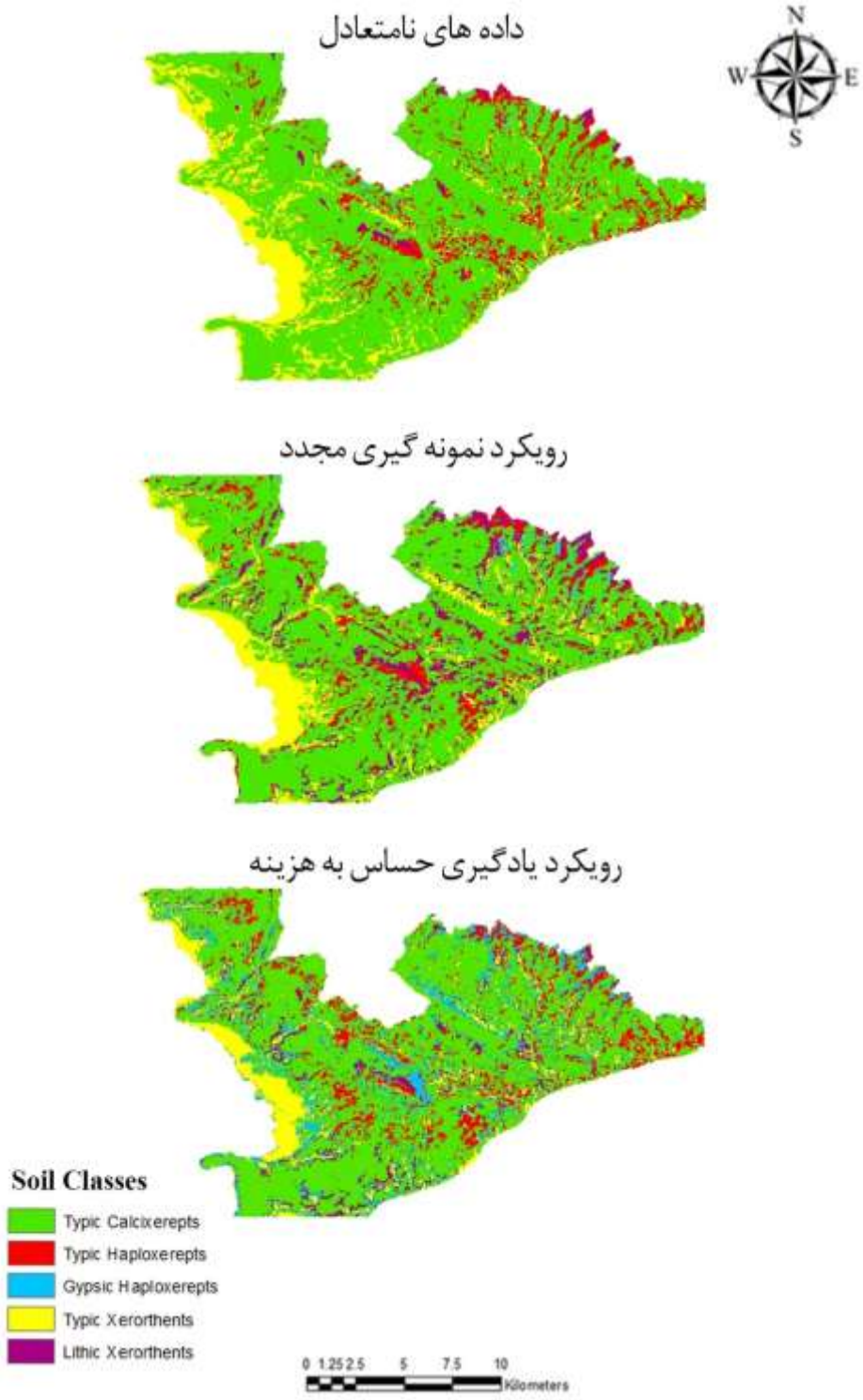
جدول ۴- صحت تولیدکننده و کاربر برای کلاس‌های خاک در سطح زیرگروه قبل و بعد از متعادل‌سازی داده‌ها (رویکرد نمونه‌گیری مجدد و یادگیری حساس به هزینه) بر اساس مدل جنگل تصادفی

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	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	100	85	A
34	71	100	25	83	50	B
100	100	NaN	75	85	0	C
34	81	65	34	100	34	D
78	100	NaN	88	91	0	E

*NaN: عدد نیست، هیچ پیش‌بینی برای این کلاس انجام‌نشده 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

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

نتیجه‌گیری

این پژوهش باهدف کارآمدی روش‌های پیش درمانی یا بهبود داده‌های نامتعادل برای تولید نقشـــه‌های خاک با دقت بالاتر و مدیریت بهتر اراضـی کشـاورزی انجام گرفت. در این مطالعه دو رویکرد یادگیری حسـاس به هزینه و نمونه‌گیری مجدد برای حل مســـئله عدم تعادل داده‌ها مورداســـتفاده قرار گرفت. نتایج نشـــان داد که روش‌های بهبود داده‌های نامتعادل می‌تولند ســـبب افزایش دقت مدل در تشـــخیص کلاس‌های اقلیت شـــده و درنتیجه منجر به افزایش دقت پیش‌بینی مکانی کلاس‌های خاک و تهیه نقشـه‌های دقیق‌تر گردد. از طرفی رویکرد یادگیری حسـاس به هزینه با تمرکز بر روی کلاس‌های با تکرار کم، نشان داد که می‌تواند به‌عنوان یک مدل برتر در دیگر مناطق نیز مورداستفاده قرار گیرد. با توجه به اینکه تحقیقات درزمینه داده‌های نامتعادل در خاک محدود اسـت، این مطالعه می‌تواند یک راه‌حل مؤثر برای مقابله با داده‌های نامتعادل در کلاس‌های خاک و تولید نقشه‌های رقومی خاک با دقت بالا باشد.

منابع

References

1. Abdi, L. and 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.
2. Breiman, L., J. H. Friedman, R. A. Olshen and Stone, C. J., 1984. Classification and Regression Trees.
3. Breiman, L., 2001. Random forests. Machine Learning, 45(1): 5-32.
4. Breiman, L., and Cutler, A., 2004. Random Forests. Department of Statistics, University of Berkeley.
5. Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A. and Edwards Jr, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 239:68-83.
6. Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 37(1): 35-46.
7. Devi, D., Biswas, S.K. and Purkayastha, B., 2019, July. A cost-sensitive weighted random forest technique for credit card fraud detection. In 2019 10th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-6). IEEE.
8. Fantappiè, M., L'Abate, G., Schillaci, C. and Costantini, E.A., 2023. Digital soil mapping of Italy to map derived soil profiles with neural networks. Geoderma Regional, 32, p.e00619.
9. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. and Herrera, F., 2018. Learning from imbalanced data sets (Vol. 10, pp. 978-3). Cham: Springer.
10. Heung, B., HO, H. C., Zhang, J., Knudby, A., Bulmer, C. E. and Schmidt, M. G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265: 62–77.

11. Jensen, J.R., 1996. Introductory digital image processing: a remote sensing perspective (No. Ed. 2). Prentice-Hall Inc.
12. Kang, M., Liu, Y., Wang, M., Li, L. and 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.
13. Khaledian, Y. and Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling, 81:401-418.
14. Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4):221-232.
15. Kuhn, M. and Johnson, K., 2013. Applied predictive modeling (Vol. 26, p. 13). New York: Springer.
16. Malone, B.P., McBratney, A.B., Minasny, B. and Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma, 154(1-2):138-152.
17. McBratney, A.B., Santos, M.M. and Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1-2):3-52.
18. Meng, X.T., Yan, F.G., Cao, B.X., Jin, M. and Zhang, Y., 2022. Efficient real-valued DOA estimation based on the trigonometry multiple angles transformation in monostatic MIMO radar. Digital Signal Processing, 123:103437.
19. 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.
20. Minasny, B. and Hartemink, A.E., 2011. Predicting soil properties in the tropics. Earth-Science Reviews, 106(1-2):52-62.
21. Minasny, B., McBratney, A.B., Malone, B.P. and Wheeler, I., 2013. Digital mapping of soil carbon. Advances in agronomy, 118:1-47.
22. 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. (In Persian with English abstract)
23. Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A. and Sharififar, A., 2021. Digital mapping of soil classes using spatial extrapolation with imbalanced data. Geoderma Regional, 26:e00422.
24. Rahimi Mashkale, M., Delavar, M.A., Jamshidi, M. and Sharififar, A., 2023. Improving the classification of Soil imbalanced data using machine learning algorithms in Some Part of Zanjan provice land. Agricultural Engineering, 46(1):61-82.(In Persian with English abstract)
25. Rahimi, M., Amirdelavar, M., Jamshidi, M. and Sharififar, A., 2023. Modeling Spatial Distribution of Soil Classes Using Machine Learning Algorithms in Some Parts of Zanjan Provice. Iranian Journal of Soil Research, 37(2):147-165. (In Persian with English abstract)
26. Schoeneberger, P. J., Wysocki, D. A., and Benham, E. C. (Eds.)., 2012. Field book for describing and sampling soils. Government Printing Office.

27. Sharififar, A., Sarmadian, F. and 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.
28. Sharififar, A., Sarmadian, F., Malone, B.P. and Minasny, B., 2019b. Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma, 350:84-92.
29. Sharififar, A. and Sarmadian, F., 2023. Coping with imbalanced data problem in digital mapping of soil classes. European Journal of Soil Science, 74(3):e13368.
30. Soil and Water Research Institute., 2010. Site Selection, Soil Survey and Land Evaluation for Development of Orchards in Zanjan Province, Iran. (In Persian with English abstract)
31. Soil science division staff. 2017. Soil survey manual". USDA Handbook 18: 120-131.
32. Soil Survey Staff. 2022. Keys to soil taxonomy, 13th edition. USDA Natural Resources Conservation Service.
33. Statistical Yearbook of Zanjan Province. 2019. Land and Climate, National Statistics Organization. (In Persian with English abstract)
34. Taghizadeh-Mehrjardi, R., Minasny, B., Toomanian, N., Zeraatpisheh, M., Amirian-Chakan, A. and Triantafilis, J., 2019. Digital mapping of soil classes using ensemble of models in Isfahan region, Iran. Soil Systems, 3(2):37.
35. Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T. and Schmidt, K., 2020. Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. Geoderma, 376:114552.
36. USDA. 2004. Soil survey laboratory methods manual. Soil survey investigations report, 42.
37. Vincent, S., Lemercier, B., Berthier, L. and Walter, C., 2018. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma, 311:130-142.
38. Wadoux, A.M.C., Minasny, B. and McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210:103359.
39. Yu, H., Sun, C., Yang, X., Zheng, S., Wang, Q. and Xi, X., 2018. LW-ELM: a fast and flexible cost-sensitive learning framework for classifying imbalanced data. IEEE Access, 6:28488-28500.
40. Zhang, C., Tan, K.C., Li, H. and 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.
41. Zhang, C., Wang, X., Liu, J., Li, M., and Zhang, J., 2021. Cost-sensitive soil mapping using a deep learning approach. ISPRS Journal of Photogrammetry and Remote Sensing, 179:172-183.
42. Zinck, J.A., Metternicht, G., Bocco, G. and Del Valle, H., 2016. Geopedology. An integration of geomorphology and pedology for soils and landscape studies: Springer International Publishing Switzerland, 556p.