Gorgan University of Agricultural Sciences and Natural Resources

Journal of Soil Management and Sustainable Production

Print ISSN: 2322-1267
Online ISSN: 2322-1275

Soil Science Society of Iran

Using of Cost-sensitive Learning Approach for Prediction of Imbalanced Soil Classes

Mastaneh Rahimi Mashkaleh[1], Mohammad Amir Delavar[*2], Mohammad Jamshidi[3]

1. Ph.D. Graduate, Dept. of Soil Science, Faculty of Agriculture, University of Zanjan, Zanjan, Iran. E-mail: mastanehrahimi@znu.ac.ir
2. Corresponding Author, Associate Prof., Dept. of Soil Science, Faculty of Agriculture, University of Zanjan, Zanjan, Iran. E-mail: amir-delavar@znu.ac.ir
3. Assistant Prof., Soil and Water Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran. E-mail: mohammadjamshidi@yahoo.com

Article Info

Article type:
Full Length Research Paper

Article history:
Received: 11.20.2023
Revised: 06.07.2024
Accepted: 06.09.2024

Keywords:
Balanced Accuracy,
Machine Learning,
Minority Class,
Random Forest

ABSTRACT

Background and Objectives: Optimal soil management and sustainable agricultural development require access to accurate and reliable information regarding soil conditions and classification. The precise prediction of soil classes and their spatial distribution is crucial. The application of machine learning methods, particularly the cost-sensitive learning approach, can improve the accuracy and efficiency of soil classification by addressing class imbalance. This study aimed to enhance soil class prediction by applying a cost-sensitive learning approach to account for imbalanced class distribution in a region in southwestern Zanjan Province, Iran.

Materials and Methods: A total of 148 soil profiles were excavated using a regular grid pattern with an average spacing of 500 meters (with some locations up to 700 meters based on expert recommendations). The profiles were described and classified through laboratory analyses up to the family level. The study incorporated various environmental covariates, including geomorphological and geological map data, a digital elevation model (DEM), and Landsat 8 satellite images. Using Principal Component Analysis (PCA) and expert knowledge, a set of key covariates-including geomorphological maps, geological data, analytical hill shading, sunrise, valley depth, LS Factor, channel network distance, topographic wetness index, and multi-resolution ridge top flatness-were selected as the most effective predictors of soil classes. The soil-landscape relationship was modeled using the Random Forest (RF) algorithm and an ensemble model after data balancing in RStudio software.

Results: The soil classes in the study area at the subgroup level were categorized into five imbalanced classes, including Typic Calcixerepts, Typic Haploxerepts, Gypsic Haploxerepts, Typic Xerorthents, and Lithic Xerorthents. The overall accuracy and Kappa coefficient for evaluating the soil map using the Random Forest model were 65% and 0.32 before data balancing and 86% and 0.77 after balancing the data using the cost-sensitive learning approach. The accuracy of soil class predictions at the subgroup level improved significantly after balancing. The two minority classes-Gypsic Haploxerepts and Lithic Xerorthents-were predicted with

100% user accuracy and 91% and 85% producer accuracy, respectively. However, the sensitivity index for these minority classes was initially zero, indicating that no correct predictions had been made before balancing. The specificity index values for Gypsic Haploxerepts and Lithic Xerorthents were 1 and 0.97, respectively, confirming that the model was highly capable of distinguishing these classes from others. The balanced accuracy values revealed that differentiating Gypsic Haploxerepts (0.50) and Lithic Xerorthents (0.49) remained more challenging compared to other classes, but the model still achieved relatively accurate predictions.

Conclusion: The findings confirm that handling imbalanced data using a cost-sensitive learning approach significantly improves the accuracy of soil class predictions and the quality of produced soil maps. By focusing on minority classes, this method reduces prediction errors and enhances model accuracy. The results demonstrate that the Random Forest algorithm, combined with the cost-sensitive learning approach, substantially improves the differentiation of soil classes, particularly minority classes, making it a valuable tool for soil classification and management.

Cite this article: Rahimi Mashkaleh, Mastaneh, Delavar, Mohammad Amir, Jamshidi, Mohammad. 2025. Using of Cost-sensitive Learning Approach for Prediction of Imbalanced Soil Classes. *Journal of Soil Management and Sustainable Production*, 14 (4), 53-73.

DOI: 10.22069/EJSMS.2025.22001.2128
Publisher: Gorgan University of Agricultural Sciences and Natural Resources

نشریه مدیریت خاک و تولید پایدار
شاپا چاپی: ۱۲۶۷-۲۳۲۲
شاپا آنلاین: ۱۲۷۵-۲۳۲۲
انجمن علوم خاک ایران

کاربرد رویکرد یادگیری حساس به هزینه برای پیش‌بینی کلاس‌های نامتعادل خاک

مستانه رحیمی مشکله[۱]، محمدامیر دلاور*[۲]، محمد جمشیدی[۳]

۱. دانش‌آموخته دکتری گروه علوم خاک، دانشکده کشاورزی، دانشگاه زنجان، زنجان، ایران. رایانامه: mastanehrahimi@znu.ac.ir
۲. نویسنده مسئول، دانشیار گروه علوم خاک، دانشکده کشاورزی، دانشگاه زنجان، زنجان، ایران. رایانامه: amir-delavar@znu.ac.ir
۳. استادیار مؤسسه تحقیقات خاک و آب، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران. رایانامه: mohammadjamshidi@yahoo.com

اطلاعات مقاله

نوع مقاله:
مقاله کامل علمی- پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۸/۲۹
تاریخ ویرایش: ۱۴۰۳/۰۳/۱۸
تاریخ پذیرش: ۱۴۰۳/۰۳/۲۰

واژه‌های کلیدی:
جنگل تصادفی،
صحت متعادل،
کلاس اقلیت،
یادگیری ماشین

چکیده

سابقه و هدف: مدیریت بهینه خاک و توسعه پایدار کشاورزی، نیاز به دسترسی اطلاعات دقیق و معتبر در مورد وضعیت و طبقه‌بندی خاک دارد و پیش‌بینی دقیق کلاس‌های خاک و تعیین مکانی آن‌ها از اهمیت بالایی برخوردار است. استفاده از روش‌های یادگیری ماشین و به‌خصوص رویکرد یادگیری حساس به هزینه می‌تواند با در نظر گرفتن نامتوازنی در توزیع کلاس‌های خاک، به بهبود دقت و کارایی پیش‌بینی کلاس‌های خاک کمک کرده و اطلاعات ارزشمندی برای مدیریت بهینه خاک و کشاورزی فراهم کند. با این هدف، این مطالعه در بخشی از اراضی جنوب غربی استان زنجان انجام شد.

مواد و روش‌ها: تعداد ۱۴۸ خاک‌رخ با روش الگوی شبکه‌بندی منظم و میانگین فاصله ۵۰۰ متر حفر، تشریح و با تجزیه‌وتحلیل آزمایشگاهی تا سطح فامیل رده‌بندی شد. 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، 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 برای پیش‌بینی کلاس‌های خاک و ورودی مدل‌ها انتخاب شد. مدل‌سازی رابطه خاک – زمین‌نما با استفاده از الگوریتم یادگیرنده جنگل تصادفی و رویکرد یادگیری حساس به هزینه در محیط نرم‌افزار "Rstudio" انجام شد.

یافته‌ها: خاک‌های منطقه در پنج کلاس با توزیع نامتعادل تا سطح زیرگروه شامل تیپیک کلسی‌زرپتز، تیپیک هاپلوزرپتز، جیپسیک هاپلوزرپتز، تیپیک زراورتنتز و لیتیک زراورتنتز بودند.

مقادیر صحت کلی و ضریب کاپا برای ارزیابی نقشه خاک در مدل جنگل تصادفی ٦٥ درصد و ۰/۳۲ و در رویکرد یادگیری حساس به هزینه ٨٦ درصد و ۰/۷۷ به دست آمد. مقادیر صحت‌سنجی پیش‌بینی کلاس‌های خاک در سطح زیرگروه نشان داد پس از متعادل‌سازی با رویکرد یادگیری حساس به هزینه تمامی کلاس‌های خاک به‌ویژه دو کلاس اقلیت جیپسیک هاپلوزرپتز و لیتیک زراورتنتز به ترتیب با مقادیر صحت کاربر ۱۰۰ درصد و صحت تولیدکننده ۹۱ و ۸۵ درصد، با صحت بسیار بالایی پیش‌بینی شدند. مقادیر شاخص حساسیت برای دو کلاس اقلیت جیپسیک هاپلوزرپتز (صفر) و لیتیک زراورتنتز (صفر) نشان می‌دهد که هیچ پیش‌بینی صحیحی برای این دو کلاس اقلیت انجام نگرفته است. مقادیر شاخص ویژگی برای کلاس‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنتز به ترتیب برابر ۱ و ۰/۹۷ بود. این مقادیر نشان می‌دهند که توانایی مدل جنگل تصادفی در تشخیص این دو کلاس نسبت به سایر کلاس‌ها بسیار بالاتر است. نتایج صحت متعادل نشان داد که بااین‌که تشخیص مدل در تمایز کلاس‌های اقلیت جیپسیک هاپلوزرپتز و لیتیک زراورتنتز با مقادیر ۰/۵۰ و ۰/٤۹ نسبت به سایر کلاس‌ها مشکل‌تر است اما بااین‌وجود مدل می‌تواند به‌صورت نسبتاً خوب کلاس‌ها را پیش‌بینی کند.

نتیجه‌گیری: نتایج مطالعه بیانگر آن است که روش بهبود داده‌های نامتعادل با رویکرد یادگیری حساس به هزینه سبب افزایش دقت پیش‌بینی در کلاس‌های خاک و نقشه تولیدشده می‌شود. تمرکز مدل در روش یادگیری حساس به هزینه‌بر روی داده‌های با فراوانی کم (اقلیت) است و این موضوع، موجب کاهش خطای پیش‌بینی و افزایش دقت مدل می‌گردد. نتایج نشان داد که الگوریتم جنگل تصادفی با استفاده از رویکرد یادگیری حساس به هزینه می‌تواند بهبود معناداری در تمایز دادن کلاس‌های خاک به‌ویژه کلاس‌های اقلیت داشته باشد.

استناد: رحیمی مشکله، مستانه، دلاور، محمدامیر، جمشیدی، محمد (۱۴۰۳). کاربرد رویکرد یادگیری حساس به هزینه برای پیش‌بینی کلاس‌های نامتعادل خاک. *نشریه مدیریت خاک و تولید پایدار*، ۱۴ (۴)، ۷۳-۵۳.

DOI: 10.22069/EJSMS.2025.22001.2128

ناشر: دانشگاه علوم کشاورزی و منابع طبیعی گرگان

مقدمه

درک دقیق از طبقه‌بندی خاک و محدودیت‌ها و شناخت شرایط و توانایی‌های مرتبط با آن، اساس توسعه پایدار و بهره‌وری بهینه از اراضی را فراهم می‌کند. بهره‌وری بهینه از اراضی، یک چالش مهم مدیریتی در سراسر جهان است که نیازمند اطلاعات دقیق مکانی در مورد خاک و اراضی است تا به استفاده مؤثر و پایدار از این منابع ارزشمند بپردازد (۱ و ۲). در این راستا نقشه‌های خاک نقش مؤثر و مهمی در تعیین راهکارها و برنامه‌های مناسب مدیریت زمین مبتنی بر شرایط و قابلیت‌های خاص انواع خاک‌ها دارند (۳). روش‌های نقشه‌برداری رقومی خاک به‌طور گسترده برای تولید نقشه‌های خاک استفاده می‌شوند و به‌عنوان ابزار عملیاتی اصلی برای تولید داده‌های خاک در مناطق بزرگ و با وضوحی معمولاً بالاتر از استفاده در نقشه‌برداری مرسوم تبدیل شده‌اند (٤ و ٥). در واقع نقشه‌برداری رقومی خاک ابزاری قدرتمند برای افزایش جزئیات مکانی اطلاعات خاک در مناطق وسیع است که برای رسیدگی به مسائل زراعی و زیست‌محیطی ضروری است (٦).

در سال‌های اخیر، علم خاکشناسی شاهد توسعه سریع در زمینه فعالیت‌های نقشه‌برداری رقومی خاک بوده است. این توسعه، ناشی از همگرایی چندین عامل مهم از جمله افزایش دسترسی به داده‌ها، سهولت دسترسی به داده‌های مکانی محیطی و توسعه راه‌حل‌های نرم‌افزاری با استفاده از ابزارهای محاسباتی برای تجزیه‌وتحلیل دقیق‌تر این داده‌ها است (۷). یکی از راه‌های مؤثر برای افزایش دانش در زمینه خاک و بهبود مدیریت منابع زمین، نقشه‌برداری رقومی کلاس‌های خاک است (۸). اکثر نقشه‌های کلاس خاک که در سطح ملی تهیه می‌شوند اگرچه مفید هستند، اما به دلیل نمایش داده‌های نامتعادل[1] در الگوی توزیع کلاس‌های خاک، همواره با چالش‌هایی مواجه هستند (۹). ازاین‌رو، لازم است که نقشه‌های خاک با وضوح و دقت بالاتر توسعه داده شوند تا قادر به ارائه نقشه‌هایی با اطلاعات مکانی دقیق برای تصمیم‌گیرندگان ارائه و بتوانند برای بهبود مدیریت اراضی و دستورالعمل‌های مدیریتی مورداستفاده قرار گیرند (۱۰).

یکی از مشکلات بسیار رایج که معمولاً خاک‌شناسان و پژوهش‌گران با آن مواجه هستند، مسئله عدم تعادل در تعداد مشاهده‌ها برای انواع مختلف خاک است. این عدم تعادل ممکن است به دلیل وجود عواملی که در فرآیند تشکیل و توسعه خاک تأثیر دارند، رخ دهد (۱۱، ۱۲ و ۱۳). این موضوع در نقشه‌برداری رقومی خاک منجر به تولید نقشه‌های با دقت بسیار پایین و حذف کلاس‌های با تعداد مشاهده‌های بسیار کم‌تر می‌شود که طبقات اقلیت[2] نامیده می‌شوند (۳). اغلب مدل‌های یادگیری ماشین و شبیه‌سازی‌ها فرض می‌کنند که داده‌ها به‌طور متعادل توزیع شده‌اند؛ بنابراین زمانی که با داده‌های نامتعادل آموزش داده می‌شوند، نتایج ضعیفی را به دنبال دارند (١٤ و ١٥). مشکل عدم تعادل داده‌ها موجب کاهش دقت و از دست دادن کلاس‌های اقلیت در پیش‌بینی‌های نهایی و تولید نقشه‌های نامطمئن یا گمراه‌کننده می‌شوند (۱۲).

در طول دهه گذشته، پژوهش‌گران با بهره‌گیری از تکنیک‌های یادگیری ماشین، به توانایی پیش‌بینی دقیق‌تر و قابل‌اعتمادتری از ویژگی‌های خاک با داده‌های محدود دست پیدا کرده‌اند. در حوزه نقشه‌برداری رقومی خاک، تکنیک‌های یادگیری ماشین به‌عنوان یک ابزار اساسی برای استخراج روابط بین ویژگی‌های خاک و متغیرهای کمکی به‌کار می‌روند (۵). یکی از کاربردهای عمده روش‌های یادگیری ماشین، شناسایی و پیش‌بینی الگوهای موجود در

1- Imbalanced data

2- Minority classes

چندین مجموعه داده بزرگ حاصل از داده‌های ماهواره‌ای یا پارامترهای مشتق‌شده از مدل رقومی ارتفاع است (١٦). این داده‌ها به‌عنوان کنترل‌کننده‌ها یا پیش‌بینی‌کننده‌های محیطی که نماینده عوامل خاک‌سازی هستند، نقش دارند (۱۷). الگوریتم‌های زیادی برای یادگیری ماشین در دسترس هستند اما از جمله رویکردهای یادگیری ماشین برای پرداختن به مسأله عدم تعادل داده‌ها که توسط محققان علوم خاک ارائه شده است می‌توان به نمونه‌گیری مجدد[1] و مدل‌های تجمعی[2] اشاره کرد (۳، ۹ و ۱۸).

رویکرد نمونه‌برداری مجدد، ممکن است راه‌حلی برای مواجهه با مشکل عدم تعادل کلاس خاک باشد، اما به دلیل از دست رفتن بخشی از داده‌ها ممکن است بهبود معنی‌داری در پیش‌بینی درست کلاس‌های اقلیت را فراهم نکند. رویکرد مدل تجمعی نیز به‌علت ترکیب چندین مدل ضعیف و قوی با همدیگر تحت‌تأثیر مدل‌های ضعیف‌تر بوده و ممکن است در پیش‌بینی مکانی برخی از کلاس‌های اقلیت خاک دقت بالایی نشان ندهد؛ بنابراین نیاز به بررسی رویکردهایی است که به‌احتمال‌زیاد حفظ کلاس اقلیت خاک در نقشه و بهبود دقت نقشه را نسبت به رویکردهای نمونه‌برداری و تجمعی، بدون ایجاد تغییری در داده‌های ورودی، تضمین کند. یکی از رویکردهایی که در این زمینه توسط پژوهش‌گران پیشنهاد می‌گردد، روش یادگیری حساس به هزینه[3] است. 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 (٦ و ۱۹). نتایج برخی از مطالعات نشان داده که همبستگی بالایی بین یادگیری حساس به هزینه و طبقه‌بندی نامتعادل وجود دارد و رویکرد مناسب‌تری نسبت به تکنیک‌های نمونه‌گیری مجدد است (۲۰، ۲۱ و ۲۲). این پژوهش به بررسی چالش‌ها و راه‌حل‌های مرتبط با پیش‌بینی مکانی کلاس‌های نامتعادل خاک با استفاده از رویکرد یادگیری حساس به هزینه در بخشی از اراضی استان زنجان در جنوب غربی ایران می‌پردازد و احتمال می‌دهد که استفاده از این رویکرد می‌تواند منجر به تحلیل دقیق‌تر و کاربردی‌تر داده‌های خاک شود. این پژوهش یک رویکرد نوآورانه است که با دقت بیش‌تری به تحلیل داده‌های خاک می‌پردازد و در نتیجه اطلاعات مفیدی را در اختیار تصمیم‌گیران و برنامه‌ریزان مرتبط با منطقه قرار می‌دهد.

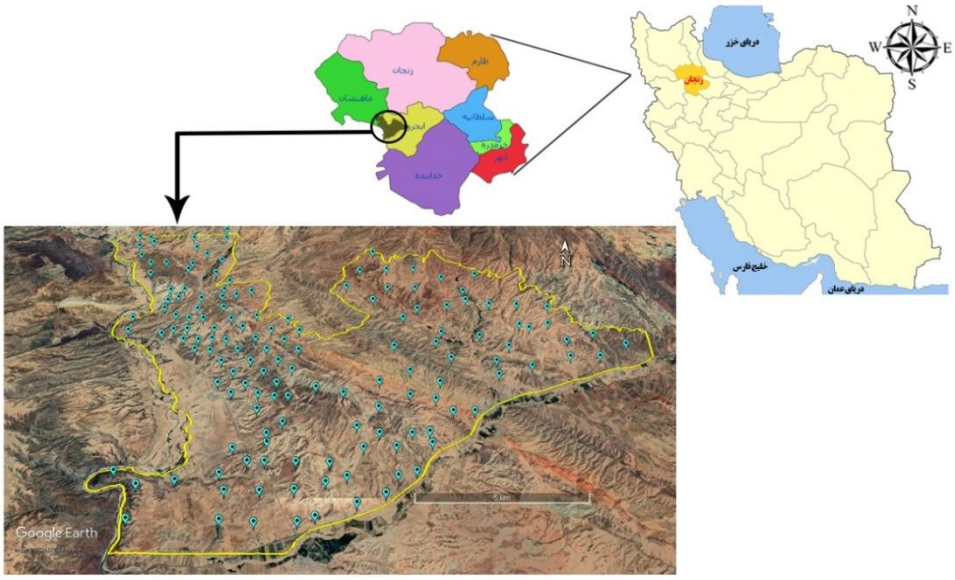
1- Resampling method
2- Ensemble models
3- Cost sensitive learning

مواد و روش‌ها

ویژگی‌های منطقه موردمطالعه: منطقه موردمطالعه در شمال غربی ایران و غرب استان زنجان حدفاصل عرض‌های جغرافیایی ′۳۱ °۳٦ تا ′۳۷ °۳٦ شمالی و طول‌های جغرافیایی ′۲۱ °٤۷ تا ′۱۱ °٤۸ شرقی با مساحت ۱۳۸۲۳ هکتار واقع‌شده است (شکل ۱). با توجه به آمار بلندمدت ۲۰ ساله (۱۳۹۸-۱۳۷۸)، متوسط بارندگی سالانه در این منطقه حدود ۳٤۰ میلی‌متر و متوسط دمای سالانه حدود ۱۳ درجه سلسیوس است. ارتفاع متوسط این منطقه از سطح دریا حدود ۱٤۸۲ متر و بین ۱۳۷۹ تا ۱٦۹۹ متر متغیر است (۲۳). خاک‌های منطقه دارای رژیم حرارتی مزیک[4] و رژیم رطوبتی زریک[5] هستند. مهم‌ترین سازندهای زمین‌شناسی منطقه مربوط به دوران پرکامبرین، پالئوزوئیک، مزوزوئیک و سنوزوئیک است و شامل چهار بخش عمده لایه‌های کربناته، سنگ‌آهک، کنگلومرا و مواد آتش‌فشانی هستند. مهم‌ترین واحدهای چشم‌انداز منطقه تپه‌ماهورها[6] و دشت‌های دامنه‌ای[7] هستند. منطقه موردمطالعه دارای

4- Mesic
5- Xeric
6- Hill lands
7- Piedmont plains

پوشش گیاهی تنک از جمله گون بوده و در زمره مراتع ضعیف قرار می‌گیرد و بخش کمی از آن شامل اراضی کشاورزی است. خاک‌های منطقه از مواد مادری حاصل از رسوبات آبرفتی یا آبرفتی – بادرفتی تشکیل شده‌اند که از مارن‌های گچی، سنگ‌آهک و ماسه‌سنگ منشأ گرفته‌اند (۲٤).



شکل ۱- موقعیت منطقه موردمطالعه و نقاط نمونه‌برداری.

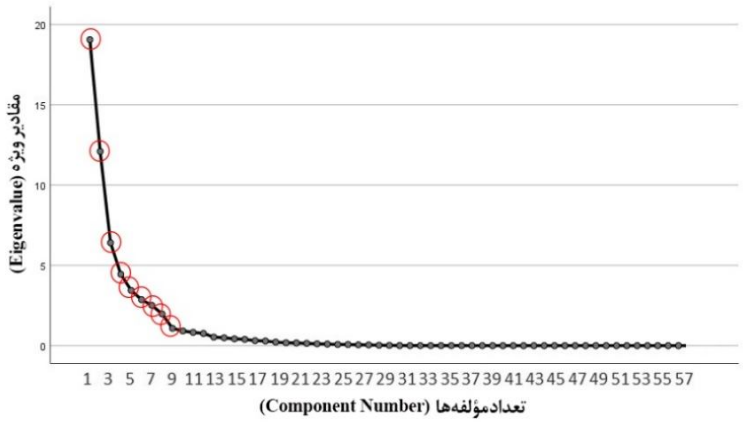
Figure 1. Location of the study area and sampling points.

نمونه‌برداری و طبقه‌بندی خاکرخ‌ها: پس از انجام بازدیدهای صحرایی، تعداد ۱٤۸ خاکرخ بر اساس یک الگوی شبکه‌ای منظم با میانگین فاصله ۵۰۰ متر (در برخی مناطق بر اساس نظر کارشناس و شرایط محلی تغییر یافتند) حفر و نمونه‌برداری از تمامی خاکرخ‌های حفرشده انجام گرفت. نمونه‌ها پس از هوا خشک شدن از الک دو میلی‌متری عبور داده شد و تجزیه‌های فیزیکوشیمیایی شامل تعیین درصد ذرات تشکیل‌دهنده خاک (۲۵)، واکنش خاک (۲٦)، کربنات کلسیم معادل خاک (۲۷)، ظرفیت تبادل کاتیونی (۲۸)، قابلیت هدایت الکتریکی (۲۹)، کربن آلی (۳۰) بر روی تمام نمونه‌ها و در صورت مشاهده صحرایی بر روی برخی نمونه‌های دارای گچ (۳۱) انجام شد. نمونه‌ها پس از هوا خشک شدن و عبور از الک دو میلی‌متری، طبق روش‌های استاندارد (۳۲) تحت تجزیه‌های فیزیکی و شیمیایی قرار گرفتند. بر اساس نتایج تشریح خاکرخ‌های حفرشده و نتایج آزمایشگاهی نمونه‌های خاک، رده‌بندی تمامی خاکرخ‌ها بر مبنای سامانه رده‌بندی آمریکایی (۳۳) تا سطح فامیل تعیین گردید.

استخراج متغیرهای کمکی: در این پژوهش از اطلاعات نقشه ژئومورفولوژی، نقشه زمین‌شناسی، داده‌های سنجش‌ازدور و اطلاعات توپوگرافی برای استخراج متغیرهای کمکی استفاده شد. برای این منظور نقشه زمین‌شناسی با مقیاس ۱:۲۵۰۰۰۰ از سازمان زمین‌شناسی کشور تهیه و در محیط سامانه اطلاعات جغرافیایی ArcGIS نسخه ۱۰/۷ زمین مرجع و رقومی‌سازی شد. ۱۸ شاخص پستی‌وبلندی با استفاده از مدل رقومی ارتفاع با قدرت تفکیک مکانی ۳۰ × ۳۰ متر از سنجنده استر، در محیط نرم‌افزار SAGA GIS (نسخه ۷/۹) استخراج شد (۳٤). شاخص‌های سنجش‌ازدوری (۳٦ متغیر) با استفاده از تصاویر

سنجنده (OLI/TIRS) ماهواره لندست ۸ با قدرت تفکیک مکانی ۳۰ × ۳۰ متر (USGS 2014) پس از اعمال تصحیح‌های رادیومتریکی و اتمسفری در محیط نرم‌افزار ENVI (نسخه ۵/۳) تهیه شد. نقشه ژئومورفولوژی بر اساس تلفیق لایه‌های اطلاعاتی واحدهای شکل زمین و مواد مادری به همراه تفسیر تصاویر ماهواره‌ای با مقیاس ۱:۵۰۰۰۰ بر اساس رویکرد سلسله‌مراتبی ارائه‌شده توسط زینک (۳۵) تهیه گردید.

انتخاب مؤثرترین متغیرهای کمکی: انتخاب مؤثرترین متغیرهای محیطی بر اساس رویکرد تحلیل مؤلفه اصلی[1] در نرم‌افزار SPSS و رتبه‌بندی اهمیت نسبی مدل یادگیری ماشین انجام گرفت. در روش تحلیل مؤلفه اصلی برای انتخاب مناسب‌ترین متغیرهای محیطی از میان ۵۷ متغیر محیطی تولیدشده، تعداد نه مؤلفه اصلی از PC۱ تا PC۹ که دارای مقادیر ارزش ویژه بالاتر از یک بودند، انتخاب شدند (شکل ۲). نتایج درصد واریانس منفرد نه مؤلفه نخست به ترتیب نشان‌دهنده مقادیر ۳۱/۱۵، ۱۱/٤، ۱۰/۷، ۹، ۷/۹۹، ۵/۹۵، ۵/۸۲، ۵/۷۵ و ۲ بودند که در مجموع ۸۹/۷٦ درصد از کل واریانس متغیرهای محیطی مورد را توجیه نمودند (جدول ۱). متغیرهای کمکی شامل اطلاعات نقشه‌های ژئومورفولوژی، اطلاعات زمین‌شناسی و ویژگی‌های مستخرج از مدل رقومی ارتفاع شامل تجزیه‌وتحلیل سایه‌اندازی تپه‌ها[2]، طلوع خورشید[3]، عمق دره[4]، شاخص طول در جهت شیب[5]، فاصله تا شبکه آبراهه[6]، شاخص رطوبتی توپوگرافی[7] و شاخص همواری بالای پشته با درجه تفکیک بالا[8] بالاترین ضریب ارزش ویژه را نشان دادند و مدل رقومی ارتفاع نیز بر اساس نظر کارشناس به فرآیند مدل‌سازی اضافه گردید. در ادامه تعداد ۱۰ متغیر محیطی پس از یکسان‌سازی مقیاس‌ها در محیط نرم‌افزار SAGA GIS به‌عنوان مؤثرترین متغیرهای محیطی برای پیش‌بینی کلاس‌های خاک و ورودی‌های مدل انتخاب شدند (۳٦).



شکل ۲- متغیرهای کمکی منتخب برای ورود به مدل‌سازی بر اساس رویکرد تحلیل مؤلفه اصلی.

Figure 2. Covariates selected to enter the modeling based on the PCA approach.

1- Principal component analysis, PCA
2- Analytical hill shading
3- Sunset
4- Valley depth
5- LS-Factor
6- Channel network distance, CND
7- Topographic wetness index, TWI
8- Multi-resolution ridge top flatness index, MRRTF

جدول ۱- مقادیر واریانس منفرد و تجمعی بر اساس رویکرد تحلیل مؤلفه اصلی.

Table 1. Individual and cumulative variance values based on the PCA approach.

مؤلفه‌ها Component	واریانس منفرد (٪) Individual variance (%)	واریانس تجمعی (٪) Cumulative variance (%)
PCA_1	31.15	31.15
PCA_2	11.4	42.55
PCA_3	10.7	53.25
PCA_4	9.00	62.25
PCA_5	7.99	70.24
PCA_6	5.95	76.19
PCA_7	5.82	82.01
PCA_8	5.75	87.76
PCA_9	2.00	89.76

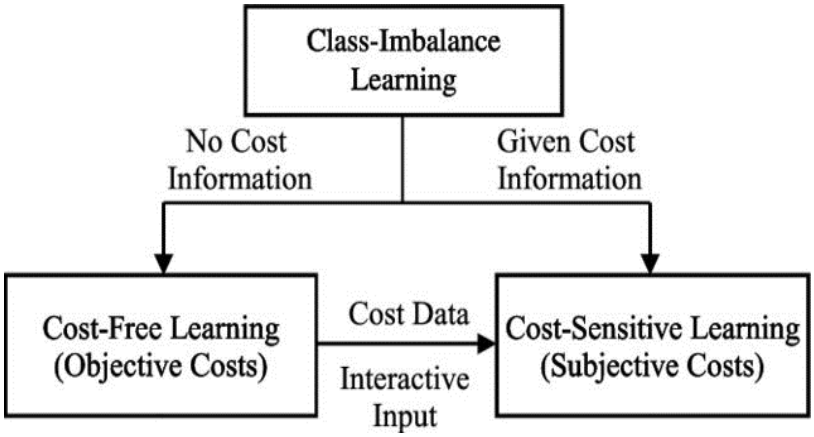
مدل‌سازی خاک و زمین‌نما بر اساس الگوریتم جنگل تصادفی: الگوریتم مورداستفاده در این مطالعه، مدل جنگل تصادفی[1] است که برای مدل‌سازی پراکنش مکانی کلاس‌های خاک در سطح زیرگروه استفاده شد. مدل جنگل تصادفی یک تکنیک یادگیرنده فعال و توسعه‌یافته از مدل طبقه‌بندی و رگرسیون درختی است. در این روش داده‌ها به‌طور تکراری برای به‌دست آوردن ارتباط بین متغیر پاسخ و متغیرهای مستقل و انجام تخمین جداسازی می‌شوند. روش جنگل تصادفی برخلاف سایر روش‌های درختی که تعداد محدودی درخت ترسیم می‌کنند، صدها یا هزاران درخت طبقه‌بندی تولید می‌کند. این روش، یک شیوه یادگیری گروهی است و برای طبقه‌بندی با ساختن تعداد درختان زیاد عمل می‌نماید (۳۷). اساس روش‌های یادگیرنده گروهی این است که گروهی از یادگیرنده‌های ضعیف، مجموعه‌ای از یادگیرنده‌های قوی را تشکیل می‌دهند. تمامی مراحل مدل‌سازی با استفاده از این روش یادگیری ماشینی با استفاده از بسته Random Forest در محیط نرم‌افزار RStudio (نسخه ۲/۳/٤۹۲) انجام گرفت. مدل جنگل تصادفی با استفاده از آنالیز حساسیت، اهمیت متغیرها در مدل‌سازی را تعیین می‌کند. این مدل با روش میانگین کاهش حداقل[2] قادر به ارائه اهمیت متغیرهای مورداستفاده در فرآیند مدل‌سازی است. در این روش مقادیر درست متغیرها با مقادیری که به‌طور تصادفی برای هر درخت تولید شده است جایگزین می‌شود و اگر این جایگزینی اثری روی خطای اندازه‌گیری نداشته باشد اهمیت آن کم و اگر مقدار خطای اندازه‌گیری افزایش یابد، آن متغیر دارای اهمیت بالایی است (۳۸).

متعادل‌سازی داده با استفاده از رویکرد یادگیری حساس به هزینه: روش‌های حساس به هزینه از جمله اولین طبقه‌بندی اصلی برای داده‌های نامتعادل است (۳۹). در واقع، یادگیری حساس به هزینه و عدم تعادل کلاس نمونه‌ها ارتباط نزدیکی با هم دارند (٤۰). طبقه‌بندی حساس به هزینه یک روش یادگیری در داده‌کاوی است که فرض می‌کند زمانی که نمونه‌ها از یک کلاس به کلاس دیگر طبقه‌بندی شوند،

1- Random forest, RF

2- Mean decrease in accuracy, MDA

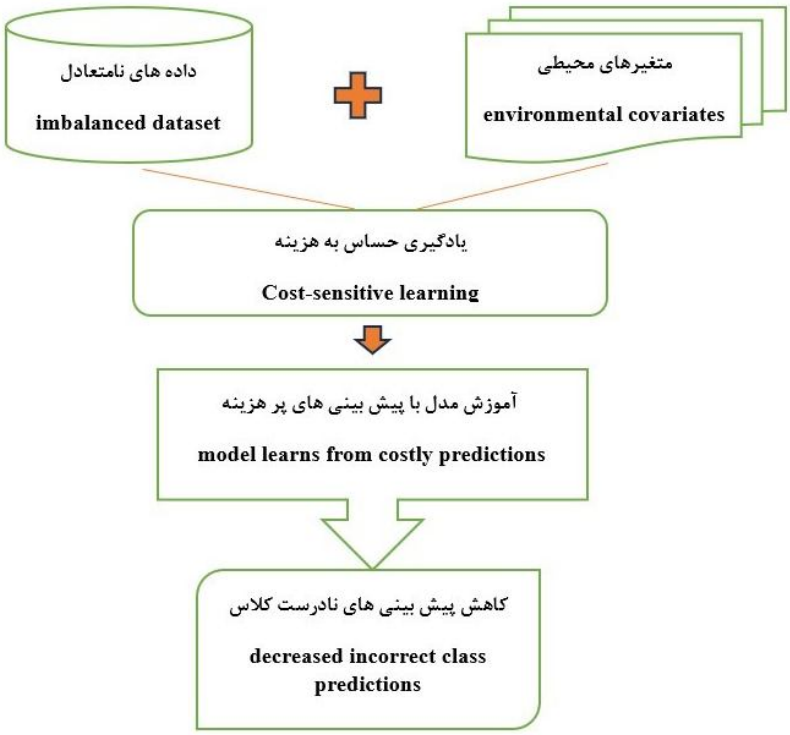
هزینه‌ها متفاوت است (شکل ۳). هدف این روش آن است که علاوه‌بر این‌که انواع مختلف هزینه‌های طبقه‌بندی نادرست را بررسی می‌کند، نحوه به حداقل رساندن هزینه کلی طبقه‌بندی نادرست را نیز موردمطالعه قرار می‌دهد (٤۱).



شکل ۳- روش یادگیری حساس به هزینه.

Figure 3. Cost-sensitive learning method.

در رویکرد حساس به هزینه، الگوریتم سعی می‌کند تا پیش‌بینی‌های کلاس را با به حداقل رساندن کل هزینه طبقه‌بندی اشتباه (به‌عنوان‌مثال، خطا) بهینه کند. یکی از مثال‌های روش حساس به توزیع کلاس در نقشه‌برداری رقومی خاک، استفاده از مدل‌های وزنی جنگل تصادفی است. جنگل تصادفی یک الگوریتم یادگیری مجموعه‌ای است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. این الگوریتم تعداد زیادی درخت تصمیم (یا جنگل) را در طول آموزش می‌سازد (٤۲). برای این‌که جنگل تصادفی استاندارد تغییر داده شود تا به هزینه حساس باشد، وزن‌هایی به کلاس‌های مختلف اختصاص داده می‌شود. برای این منظور از معکوس توزیع کلاس استفاده می‌شود. در نتیجه الگوریتم مجبور می‌شود تا بیش‌تر روی کلاس اقلیت تمرکز کند (٤۳). در این روش، مدل جنگل تصادفی به‌گونه‌ای تغییر داده می‌شود که توزیع کلاس در داده‌های آموزشی مدنظر قرار گیرد. به‌طور خاص، مدل به هر نمونه در داده‌های آموزشی، وزنی اختصاص می‌دهد که براساس معکوس فراوانی کلاس مربوطه آن نمونه محاسبه می‌شود. این باعث می‌شود که نمونه‌هایی از کلاس‌های کم‌رخداد (اقلیت)، وزن بالاتری در پیش‌بینی مدل داشته باشند و تأثیر آن‌ها در پیش‌بینی‌های مدل بیش‌تر شود. فلوچارت متعادل‌سازی داده با استفاده از روش یادگیری حساس به هزینه در شکل ٤ نشان داده شده است. در این مطالعه از تابع rf موجود در بسته Random Forest در نرم‌افزار Rstudio استفاده شد.



شکل ٤- فلوچارت متعادل‌سازی داده با استفاده از روش یادگیری حساس به هزینه.

Figure 4. Flowchart of data balancing using cost-sensitive learning method.

در رویکرد یادگیری حساس به هزینه معیاری تحت عنوان صحت متعادل[1] برآورد می‌شود. صحت متعادل یک معیار مهم در زمینه ارزیابی عملکرد مدل‌های طبقه‌بندی در مواجهه با داده‌های نامتعادل است. زمانی که در مجموعه داده‌ها تعداد نمونه‌های نامتعادلی از کلاس‌های مختلف وجود داشته باشد، به‌ویژه در مواردی که یک کلاس نسبت به دیگر کلاس‌ها تعداد نمونه‌های بسیار کم‌تری دارد، استفاده از صحت متعادل بسیار مفید است. این شاخص میانگینی از حساسیت[2] (نرخ مثبت واقعی) و ویژگی[3] (نرخ منفی واقعی) است (٤٤). رابطه ١ نحوه محاسبه صحت متعادل را نشان می‌دهد:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{Sensitivity}+\text{Specificity}}{2}\right) \quad (١)$$

بر اساس این رابطه، حساسیت نشان‌دهنده توانایی مدل در تشخیص نمونه‌های مثبت (نمونه‌های واقعی کلاس اقلیت) است (رابطه ٢) و ویژگی نشان‌دهنده توانایی مدل در تشخیص نمونه‌های منفی یا نمونه‌های واقعی کلاس اکثریت است (رابطه ٣).

$$Sensitivity = \left(\frac{\text{TP}}{\text{TP+FN}}\right) \quad (٢)$$

در رابطه اخیر، TP تعداد نمونه‌های واقعی مثبت[4] و FN تعداد نمونه‌های واقعی مثبتی است که توسط مدل به‌اشتباه تشخیص داده نشده[5] است.

$$Specificity = \left(\frac{TN}{TN+FP}\right) \quad (٣)$$

که در این رابطه، TN تعداد نمونه‌های واقعی منفی[6] و FP تعداد نمونه‌های واقعی منفی است که توسط مدل به‌اشتباه تشخیص داده نشده[7] است.

1- Balanced accuracy
2- Sensitivity
3- Specificity

4- True positive rate, TP
5- False negative rate, FN
6- True negative rate, TN
7- False positive rate, FP

صحت متعادل به ما نشان می‌دهد که مدل چقدر موفق به ایجاد تعادل بین صحت در تشخیص هر دو نوع نمونه (مثبت و منفی) در مقایسه با تعداد واقعی این نمونه‌ها است. این شاخص مقداری بین صفر و یک دارد که هرچه به عدد یک نزدیک‌تر باشد صحت پیش‌بینی برای آن کلاس خوب و در بدترین حالت صفر است که هیچ پیش‌بینی برای کلاس انجام نگرفته است (٤٤).

ارزیابی دقت مدل‌سازی: به‌منظور آموزش مدل‌ها، مجموعه خاک‌رخ‌ها (متغیرهای محیطی و کلاس‌های خاک) به‌صورت تصادفی به دو مجموعه با نسبت چهار به یک تقسیم شدند. به‌عبارت‌دیگر، ۸۰ درصد داده‌ها برای آموزش مدل و ۲۰ درصد دیگر به‌عنوان داده‌های اعتبارسنجی برای ارزیابی مورداستفاده قرار گرفتند. برای ارزیابی مدل‌ها از شاخص‌های صحت کلی نقشه[1]، صحت تولیدکننده[2]، صحت کاربر[3] و ضریب کاپا[4] استفاده شد (٤٥، ٤٦ و ٤٧).

نتایج و بحث

با توجه به این‌که شرایط ژئومورفولوژیک منطقه مطالعاتی بر روی تعدادی از خصوصیات خاک از جمله بافت خاک، عمق، درصد سنگریزه، میزان ماده آلی، میزان تجمع گچ و آهک در خاک‌رخ‌ها بیش‌ترین تأثیر را داشت موجب تمایز خاک‌های منطقه شده است (جدول ۲).

بر اساس اطلاعات جدول ۲ خاک‌ها در دو رده انتی‌سولز[5] و اینسپتی‌سولز[6] طبقه‌بندی شدند. این طبقه‌بندی در سطح زیرگروه شامل کلاس‌های تیپیک کلسی‌زرپتز[7]، تیپیک هاپلوزرپتز[8]، جیپسیک هاپلوزرپتز[9]، تیپیک زراورتنتز[10] و لیتیک زراورتنتز[11] هستند (جدول ۳). نتایج نشان داد که زیرگروه‌های خاک جیپسیک هاپلوزرپتز و لیتیک هاپلوزرپتز به‌ترتیب با فراوانی‌های ۸/۱ و ۷/٤۳ درصد به‌عنوان کلاس‌های اقلیت شناسایی شده‌اند، درحالی‌که گروه‌های خاک تیپیک کلسی‌زرپتز، تیپیک هاپلوزرپتز و تیپیک زراورتنتز به‌ترتیب با فراوانی‌های بیش‌تر از ٤٥/۹٤، ۱۷/٥٦ و ۲۰/۹٤ درصد از کل مشاهده‌ها به‌عنوان کلاس‌های اکثریت در نظر گرفته شدند.

نتایج مقادیر صحت‌سنجی پیش‌بینی مکانی هر یک از کلاس‌های خاک در شرایط معمول و بعد از رفع محدودیت داده‌های نامتعادل با رویکرد یادگیری حساس به هزینه با استفاده از الگوریتم جنگل تصادفی بر اساس دو شاخص صحت کلی و شاخص کاپا در جدول ٤ نشان داده شده است. با توجه به این نتایج مقادیر شاخص کاپا و صحت کلی قبل از متعادل‌سازی داده‌ها، به‌ترتیب برابر با ۰/۳۲ و ٦٥ درصد، پس از متعادل‌سازی داده‌ها با رویکرد یادگیری حساس به هزینه برای مدل جنگل تصادفی مقادیر ۰/۷۷ برای شاخص کاپا و ۸٦ درصد شاخص صحت کلی را نشان داد. مطابق با نتایج ارائه‌شده، روش یادگیری حساس به هزینه در پیش‌بینی مکانی کلاس‌های خاک با افزایش ۲۱ درصدی در صحت کلی و بیش از دو برابر در ضریب کاپا دقت بالاتری نسبت به قبل از بهبود داده‌ها نشان داده است. این نتایج نشان‌دهنده آن است که روش بهبود داده‌های نامتعادل با رویکرد یادگیری حساس به هزینه سبب افزایش دقت پیش‌بینی در کلاس‌های خاک و نقشه تولیدشده می‌شود. به‌عبارت‌دیگر، در روش یادگیری حساس به هزینه، تمرکز مدل بر روی داده‌های با فراوانی کم (اقلیت) است و این موضوع، موجب کاهش خطای پیش‌بینی و افزایش دقت مدل می‌گردد.

1- Overall accuracy, OA
2- Producer accuracy, PA
3- Users accuracy, UA
4- Kappa index
5- Entisols
6- Inceptisols
7- Typic Calcixerepts
8- Typic Haploxerepts
9- Gypsic Haploxerepts
10- Typic Xerorthents
11- Lithic Xerorthents

مینیه و سان (۲۰۲۱) در مطالعه خود عملکرد روش‌های یادگیری حساس به هزینه با استفاده از داده‌های نامتعادل پزشکی را مورد ارزیابی قرار دادند. آن‌ها چهار الگوریتم شامل رگرسیون لجستیک حساس به هزینه[1]، درخت تصمیم حساس به هزینه[2]، گرادیان تصادفی تقویت‌شده حساس به هزینه[3] و جنگل تصادفی حساس به هزینه را مقایسه کردند و دریافتند که الگوریتم جنگل تصادفی حساس به هزینه عملکرد و دقت بالاتری برای بهبود داده‌های نامتعادل در مقایسه با سایر الگوریتم‌ها داشته است (۲۱). کنگ و همکاران (۲۰۲۲) و دوی و همکاران (۲۰۱۹) نیز در مطالعاتی جداگانه از الگوریتم جنگل تصادفی با یادگیری حساس به هزینه استفاده کردند و دریافتند این روش تا ۳۰ درصد سبب افزایش مقادیر صحت‌سنجی نتایج شده و عملکرد و کارایی خوبی در مقایسه با روش‌های معمول جنگل تصادفی دارد و برای مجموعه‌های نمونه کوچک نیز قابل‌استفاده است (۴۸ و ۴۹).

جدول ۲- خلاصه ویژگی‌های فیزیکوشیمیایی خاک‌های موردمطالعه.

Table 2. Summary of physicochemical properties of studied soils.

گچ Gypsum	آهک Calcium carbonate	کربن آلی Organic carbon	ظرفیت تبادل کاتیونی CEC ($Cmol+kg^{-1}$)	قابلیت هدایت الکتریکی EC (dSm^{-1})	واکنش خاک pH	درصد نسبی ذرات Relative percentage of particles (%)			رنگ خاک Soil color	عمق Depth (cm)	افق Horizon
	(%)					رس Clay	سیلت Silt	شن Sand			
خاک‌رخ شماره ۱ (Profile number 1)											
-	16.7	0.46	12.90	0.49	7.98	16	28	56	10YR4/4	0-15	A
-	26.1	0.26	18.90	0.29	8.30	28	22	50	10YR5/4	15-45	Bk1
-	27.9	0.06	16.30	0.29	8.13	20	16	64	10YR5/4	45-65	Bk2
-	25.8	0.07	6.40	0.44	8.19	14	12	74	10YR5/4	65-150	C
خاک‌رخ شماره ۲ (Profile number 2)											
7	21.4	1.015	12.50	2.8	۷/٦٥	10	36	54	10YR7/2	0-15	A
23	16.9	0.61	9.60	2.7	7.86	10	36	54	10YR8/2	15-30	Bky1
10	18.9	0.39	15.60	2.9	7.81	18	36	46	10YR5/4	30-47	Bky2
20	18.3	0.35	10.60	9.26	7.62	18	30	52	7.5YR5/4	47-75	Cky
39	13.1	0.19	10.30	8.10	7.62	12	30	58	7.5YR5/4	75-150	Cy
خاک‌رخ شماره ۳ (Profile number 3)											
-	19.3	0.11	25.60	0.43	7.94	36	48	16	10YR5/6	0-10	A
-	23.3	0.13	23.90	1.04	8.77	40	48	12	10YR5/6	10-40	C
-	25.5	0.06	19.2	2.53	8.26	34	50	16		40-80	Cr
خاک‌رخ شماره ٤ (Profile number 4)											
-	13.1	0.28	9.10	0.60	7.81	10	22	68	10YR6/3	0-20	A
-	14.3	0.23	10.60	1.01	7.79	12	20	68	10YR5/3	20-45	C1
-	11.5	0.20	9.90	1.03	7.93	12	20	68	10YR5/3	45-80	C2
-	12.6	0.10	-	0.65	8.15	6	14	80	10YR5/3	80-150	C3
خاک‌رخ شماره ٥ (Profile number 5)											
-	21.6	0.25	18.00	0.76	8.02	28	51	21	10YR4/4	0-25	A
-	22.6	0.16	28.50	1.53	8.63	46	42	12	10YR4/3	25-80	Bw1
-	18.7	0.17	-	3.77	8.66	50	42	8	10YR4/3	80-150	Bw2

1- Cost sensitive logistic regression, CSLR
2- Cost sensitive decision tree, CSDT
3- Cost sensitive extreme gradient boosting, CSEGB

جدول ۳- رده‌بندی خاک‌ها بر اساس سامانه رده‌بندی خاک‌ها.

Table 3. Classification of soils based on Soil Taxonomy.

رده خاک Soil order	زیرگروه Subgroup	فامیل خاک Family soil class	تعداد خاکرخ‌ها Number of soil profiles
Inceptisols	Typic Calcixerepts	Loamy-skeletal, mixed, mesic Typic Calcixerepts	65
Inceptisols	Typic Haploxerepts	Fine-loamy, mixed, mesic Typic Haploxerepts	26
Inceptisols	Gypsic Haploxerepts	Fine, mixed, mesic Gypsic Haploxerepts	12
Entisols	Typic Xerorthents	Loamy-skeletal, mixed, calcareous, mesic Typic Xerorthents	31
Entisols	Lithic Xerorthents	Fine, mixed, calcareous, mesic Lithic Xerorthents	11

جدول ٤- صحت پیش‌بینی سطح رده‌بندی زیرگروه، قبل و بعد از متعادل‌سازی داده‌ها توسط الگوریتم جنگل تصادفی.

Table 4. The prediction accuracy of the taxonomic level of the subgroup before and after balancing the data by random forest algorithm.

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

جدول ۵ نتایج دو شاخص صحت کاربر و صحت تولیدکننده برای کلاس‌های خاک در شرایط معمول و بعد از رفع محدودیت داده‌های نامتعادل با رویکرد یادگیری حساس به هزینه در سطح زیرگروه با الگوریتم جنگل تصادفی را نشان می‌دهد. مطابق با نتایج اعتبارسنجی، در رویکرد یادگیری حساس به هزینه، پیش‌بینی مکانی تمامی زیرگروه‌های خاک با دقت نسبتاً بالاتری انجام‌گرفته است. زیرگروه‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنز که جزء کلاس‌های اقلیت محسوب می‌شوند هنگام استفاده از کلاس‌های نامتعادل توسط الگوریتم جنگل تصادفی پیش‌بینی‌نشده حذف‌شده بودند و پس از بهبود داده‌ها با رویکرد یادگیری حساس به هزینه سبب افزایش دقت پیش‌بینی مکانی کلاس‌های خاک در زیرگروه‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنز شد. نتایج نشان می‌دهد 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[1]) بهینه‌شده است. نتایج مطالعه فرناندز و همکاران (۲۰۰۹) نشان داد که روش‌های یادگیری حساس به هزینه پتانسیل بالایی برای مقابله با مشکل عدم تعادل کلاس‌بندی در داده‌کاوی و یادگیری ماشین دارند (۵۰). نتایج پژوهش‌های پژوهش‌گران در دنیا نیز بیانگر آن است که روش‌های حساس به هزینه در

1- Underfitting

نقشه‌برداری رقومی خاک مهم هستند، زیرا کمک می‌کنند تا اطمینان حاصل شود که تمام طبقات خاک بدون توجه به فراوانی آن‌ها در داده‌های آموزشی، به‌طور دقیق در نقشه‌های حاصل نمایش داده می‌شوند. با در نظر گرفتن توزیع کلاس‌های خاک در داده‌های آموزشی، این روش‌ها می‌توانند دقت و قابلیت اطمینان مدل‌های نقشه‌برداری خاک را بهبود بخشند که برای مدیریت مؤثر زمین و تصمیم‌گیری بسیار مهم است (۴۳، ۵۱، ۵۲ و ۵۳).

جدول ۵- صحت تولیدکننده و کاربر برای کلاس‌های خاک در سطح زیرگروه قبل و بعد از متعادل‌سازی داده‌ها بر اساس مدل جنگل تصادفی.

Table 5. Producer and user accuracy for soil classes at the subgroup level before and after balancing the data based on the random forest model.

صحت تولیدکننده (٪) Producer accuracy (%)		صحت کاربر (٪) User accuracy (%)		قابلیت اطمینان Reliability
داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	مدل‌های یادگیری ماشین Machine learning models
100	85	95	61	تیپیک کلسی‌زرپتز Typic Calcixerepts
83	50	71	100	تیپیک هاپلوزرپتز Typic Haploxerepts
85	0	100	NaN	جیپسیک هاپلوزرپتز Gypsic Haploxerepts
100	34	81	65	تیپیک زراورتنتز Typic Xerorthents
91	0	100	NaN	لیتیک زراورتنتز Lithic Xerorthents

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

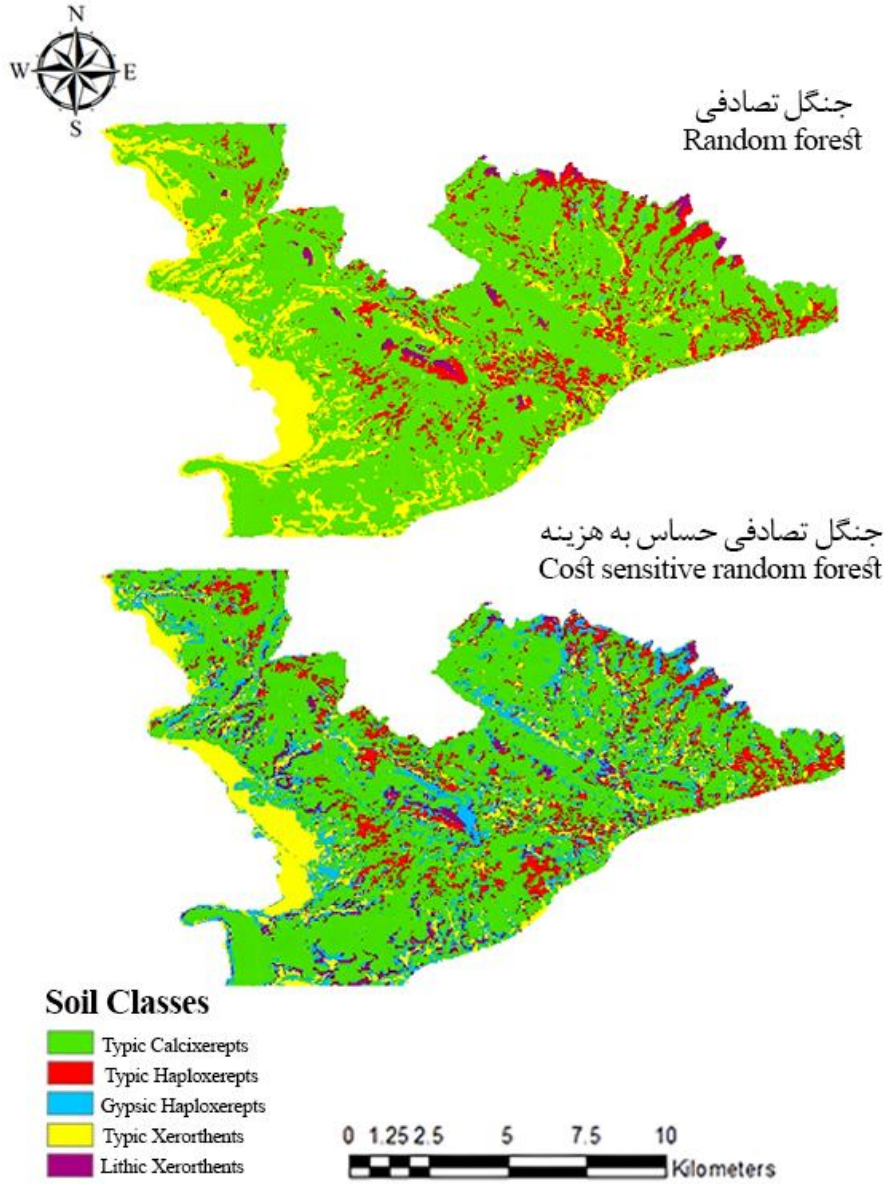
نتایج شاخص‌های حساسیت، ویژگی و صحت متعادل با رویکرد یادگیری حساس به هزینه برای الگوریتم جنگل تصادفی در جدول ٦ نشان داده شده است. مقادیر شاخص حساسیت برای دو کلاس اقلیت جیپسیک هاپلوزرپتز و لیتیک زراورتنتز نشان می‌دهد که هیچ پیش‌بینی صحیحی برای این دو کلاس اقلیت انجام نگرفته است. مقادیر شاخص ویژگی برای کلاس‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنتز به‌ترتیب برابر ۱ و ۰/۹۷ است که این مقادیر نشان می‌دهد توانایی مدل در تشخیص این دو کلاس نسبت به سایر کلاس‌ها بسیار بالاتر است. نتایج صحت متعادل، یک ترکیب از شاخص‌های حساسیت و ویژگی است و نشان می‌دهد بااین‌که که صحت مدل در تمایز کلاس‌های اقلیت جیپسیک هاپلوزرپتز و لیتیک زراورتنتز با مقادیر به‌ترتیب ۰/۵۰ و ۰/٤۹ نسبت به سایر کلاس‌ها مشکل‌تر است، اما مدل می‌تواند به‌صورت نسبتاً خوبی این کلاس‌ها را پیش‌بینی کند.

در شکل ۵ نقشه تولیدشده قبل و بعد از متعادل‌سازی داده‌ها با رویکرد یادگیری حساس به هزینه با استفاده از الگوریتم جنگل تصادفی نشان داده شده است. همان‌طور که مشاهده می‌شود پیش‌بینی مکانی برای دو کلاس اقلیت جیپسیک هاپلوزرپتز (رنگ آبی در نقشه) و لیتیک زراورتنتز (رنگ بنفش در نقشه) با دقت قابل‌قبولی انجام‌گرفته است درصورتی‌که پیش‌بینی مکانی کلاس‌های خاک با داده‌های نامتعادل منجر به حذف این دو کلاس اقلیت شده بود. ژانگ و همکاران (۲۰۱۹) معتقدند رویکردهای حساس به هزینه در مواجه با مشکلات داده‌های نامتعادل عملکرد بهتری دارند. شریفی‌فر و همکاران (۲۰۲۳) در مطالعه‌ای به مقایسه روش‌های متعادل‌سازی داده‌ها با رویکردهای مختلف نمونه‌گیری مجدد، گرادیان تصادفی تقویت‌شده، یادگیری حساس به هزینه و طبقه‌بندی تک کلاسه پرداختند و به این نتیجه رسیدند روش‌های گرادیان تصادفی تقویت‌شده و یادگیری حساس به هزینه می‌توانند به‌طور قابل‌توجهی دقت پیش‌بینی کلاس‌های خاک را برای نقشه‌برداری افزایش دهد. چنین مدل‌هایی می‌توانند حفظ طبقات اقلیت را در نقشه نهایی تضمین کنند (۸). نتایج مطالعات پژوهش‌گران مختلفی مانند ونگ و همکاران (۲۰۲۰)، فان و همکاران (۲۰۱۹) و یو و همکاران (۲۰۱۸) نیز نشان داد که روش‌های مختلف یادگیری حساس به هزینه در رسیدگی به مشکلات عدم تعادل کلاس خاک بهتر عمل می‌کنند (۵۳، ٥٤ و ۵۵).

جدول ٦- صحت پیش‌بینی سطح رده‌بندی زیرگروه قبل و بعد از متعادل‌سازی داده‌ها توسط الگوریتم جنگل تصادفی.

Table 6. Prediction accuracy of the taxonomic level of the subgroup before and after balancing the data by random forest algorithm.

پیش‌بینی کلاس‌های خاک Prediction of soil classes					
شاخص‌های صحت‌سنجی Validation indices	تیپیک کلسی‌زرپتز Typic Calcixerepts	تیپیک هاپلوزرپتز Typic Haploxerepts	جیپسیک هاپلوزرپتز Gypsic Haploxerepts	تیپیک زراورتنتز Typic Xerorthents	لیتیک زراورتنتز Lithic Xerorthents
حساسیت Sensitivity	1	0.83	0	1	0
ویژگی Specificity	0.95	0.94	1	0.94	0.97
صحت متعادل Balanced accuracy	0.97	0.89	0.50	0.98	0.49



شکل ۵- نقشه‌های تولیدشده توسط الگوریتم‌های یادگیری ماشین قبل و بعد متعادل‌سازی داده‌ها با رویکرد یادگیری حساس به هزینه.

Figure 5. Maps produced by machine learning algorithms before and after data balancing with a cost-sensitives learning approach.

نتیجه‌گیری کلی

استفاده از الگوریتم‌های معمول برای نقشه‌برداری رقومی کلاس‌های خاک، بدون توجه به تعداد نامتعادل کلاس‌های خاک در یک منطقه، می‌تواند منجر به نادیده گرفتن مدل از کلاس‌های کم‌تکرار یا از دست دادن کلاس‌های اقلیت شود و تخمین بیش‌تری از کلاس‌های اکثریت یا کلاس‌های پرتکرار داشته باشد. نتایج ارزیابی الگوریتم‌های یادگیری ماشین مختلف در پیش‌بینی کلاس‌های خاک با استفاده از رویکرد یادگیری حساس به هزینه نشان داد که توزیع نامتعادل کلاس‌های خاک می‌تواند بر خروجی مدل‌های پیش‌بینی تأثیر داشته باشد. نتایج این مطالعه نشان داد که الگوریتم جنگل تصادفی با استفاده از رویکرد یادگیری حساس به هزینه می‌تواند بهبود معناداری در

تمایز دادن کلاس‌های خاک، به‌ویژه کلاس‌های اقلیت داشته باشد و تعادل داده‌ها می‌تواند دقت پیش‌بینی مکانی کلاس‌های خاک و نقشه‌های تولیدشده را بهبود ببخشد. این نتیجه نشان می‌دهد که رفع مشکل عدم تعادل کلاس‌ها یک گام اساسی برای افزایش عملکرد مدل‌های طبقه‌بندی خاک است. با تأکید بر این نکته که مطالعات در زمینه داده‌های نامتعادل در خاک بسیار محدود هستند، این پژوهش می‌تواند به‌عنوان یک الگوی مفید برای انجام پژوهش‌ها در حوزه نقشه‌برداری رقومی خاک کمک کرده و به‌عنوان یک منبع ایده‌آل برای ایجاد شتاب در پژوهش‌های مرتبط با نقشه‌برداری رقومی خاک با توجه به تعداد نامتعادل کلاس‌ها مورداستفاده قرار گیرد و به بهبود دقت مدل‌ها و تفسیر درست‌تر انواع خاک‌ها در یک منطقه کمک نماید.

منابع

1. Garg, K. K., Anantha, K. H., Nune, R., Akuraju, V. R., Singh, P., Gumma, M. K., ... & Ragab, R. (2020). Impact of land use changes and management practices on groundwater resources in Kolar district, Southern India. *Journal of Hydrology: Regional Studies*, 31, 100732. **doi.org/ 10.1016/j.ejrh.2020.100732.**
2. Bouma, J., Bonfante, A., Basile, A., van Tol, J., Hack-ten Broeke, M. J. D., Mulder, M., ... & Hirmas, D. R. (2022). How can pedology and soil classification contribute towards sustainable development as a data source and information carrier?. *Geoderma*, 424, 115988. **doi.org/10. 1016/j.geoderma.2022.115988.**
3. 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.**
4. Lagacherie, P., Arrouays, D., & Walter, C. (2013). Cartographie numérique des sols: principe, mise en œuvre et potentialités. *Etude et Gestion des Sols*, 20 (1), 83-98.
5. Wadoux, A. M. C., Brus, D. J., & Heuvelink, G. B. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma*, 355, 113913. **doi.org/10.1016/j.geoderma.2019.113913.**
6. Vincent, S., Lemercier, 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.**
7. 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.**
8. 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.**
9. Rahimi mashkale, M., Delavar, M. A., jamshidi, M., & sharififar, A. (2023). Improving the classification of Soil imbalanced data using machine learning algorithms in Some Part of Zanjan provice 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]
10. Helfenstein, A., Mulder, V. L., Heuvelink, G. B., & Okx, J. P. (2022). Tier 4 maps of soil pH at 25 m resolution for the Netherlands. *Geoderma*, 410, 115659. **doi.org/10. 1016/j.geoderma.2021.115659.**
11. 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. **doi.org/10.1016/j.geoderma.2015.11.014.**

12. 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.**
13. 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.**
14. Zhu, B., Baesens, B., & vanden Broucke, S. K. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information sciences*, 408, 84-99. **doi.org/10.1016/j.ins.2017.04.015.**
15. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with application*s, 73, 220-239. **doi.org/10.1016/j.eswa.2016.12.035.**
16. Padarian, J., Minasny, B., & McBratney, A. B. (2019). Machine learning and soil sciences: A review aided by machine learning tools. **doi:10.5194/soil-6-35-2020.**
17. Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *Peer J*. 6, e5518. **doi: 10.7717/peerj.5518.**
18. 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. **doi:10.3390/soilsystems3020037.**
19. Jing, X. Y., Zhang, X., Zhu, X., Wu, F., You, X., Gao, Y., ... & Yang, J. Y. (2019). Multiset feature learning for highly imbalanced data classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (1), 139-156. **doi:10.1109/TPAMI.2019.2929166.**
20. Zhang, C., Tan, K. C., Li, H., & Hong, G. S. (2019). 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.**
21. Mienye, I. D., & Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25, 100690. **doi.org/10.1016/j.imu.2021.100690.**
22. Ma, Y., Zhao, K., Wang, Q., & Tian, Y. (2020). Incremental cost-sensitive support vector machine with linear-exponential loss. *IEEE Access*, 8, 149899-149914. **doi:10.1109/ACCESS. 2020.3015954.**
23. Statistical Yearbook of Zanjan Province. (2019). Land and Climate, National Statistics Organization. [In Persian]
24. Soil and Water Research Institute. (2010). Site Selection, Soil Survey and Land Evaluation for Development of Orchards in Zanjan Province, Iran. [In Persian]
25. Bouyoucos, G. J. (1962). Hydrometer method improved for making particle size analyses of soils 1. *Agronomy journal*, 54 (5), 464-465. **doi:10.2134/agronj1962.00021962005400050028x.**
26. Perry Jr, C. R., & Lautenschlager, L. F. (1984). Functional equivalence of spectral vegetation indices. *Remote sensing of environment*, 14 (1-3), 169-182. **doi:10.1016/0034-4257(84)90013-0.**
27. Lanyon, L. E., & Heald, W. R. (1983). Magnesium, calcium, strontium, and barium. Methods of Soil Analysis: Part 2 Chemical and Microbiological Properties, 9, 247-262. **doi:10.2134/agronmonogr9.2.2ed.c14.**

28. Sumner, M. E., & Miller, W. P. (1996). Cation exchange capacity and exchange coefficients. Methods of soil analysis: Part 3 Chemical methods, 5, 1201-1229. **doi:10.2136/sssabookser5.3.c40.**
29. Richards, L. A. (Ed.). (1954). Diagnosis and improvement of saline and alkali soils (No. 60). US Government Printing Office. **doi:10.1097/00010694-195408000-00012.**
30. Walkley, A., & Black, I. A. (1934). An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil science*, 37 (1), 29-38. **doi:10.1097/00010694-193401000-00003.**
31. Artieda, O., Herrero, J., & Drohan, P. J., (2006). Refinement of the differential water loss method for gypsum determination in soils. *Soil Science Society of America Journal*, 70 (6), 1932-1935. **doi:10.2136/sssaj2006.0043N.**
32. Soil Survey Staff. (2022). Keys to soil taxonomy, 13th edition. USDA Natural Resources Conservation Service.
33. Olaya, V. I. C. T. O. R. (2004). A gentle introduction to SAGA GIS. The SAGA User Group eV, Gottingen, Germany, 208.
34. 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. **doi:10.1007/978-3-319-19159-1.**
35. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer. **doi:10.1007/978-1-4614-6849-3.**
36. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32. **doi:10.1023/A:1010933404324.**
37. Breiman, L., & Cutler, A. (2004). Random Forests. Department of Statistics, University of Berkeley. **doi:10.1214/10-AOAS427.**
38. Zhao, P., Zhang, Y., Wu, M., Hoi, S. C., Tan, M., & Huang, J. (2018). Adaptive cost-sensitive online classification. *IEEE Transactions on Knowledge and Data Engineering*, 31 (2), 214-228. **doi:10.1109/TKDE.2018.2826011.**
39. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21 (9), 1263-1284. **doi:10.1109/TKDE.2008.239.**
40. Moepya, S. O., Akhoury, S. S., & Nelwamondo, F. V. (2014, December). Applying cost-sensitive classification for financial fraud detection under high class-imbalance. In *2014 IEEE international conference on data mining workshop* (pp. 183-192). IEEE. **doi:10.1109/ICDMW.2014.141.**
41. Jin, C., & Jin, S. W. (2018). Content-based image retrieval model based on cost sensitive learning. *Journal of Visual Communication and Image Representation*, 55, 720-728. **doi:10.1016/j.jvcir.2018.08.009.**
42. Zhang, J., Schmidt, M. G., Heung, B., Bulmer, C. E., & Knudby, A. (2022). Using an ensemble learning approach in digital soil mapping of soil pH for the Thompson-Okanagan region of British Columbia. *Canadian Journal of Soil Science*, 102 (03), 579-596. **doi:10.1139/cjss-2021-0091.**
43. Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The balanced accuracy and its posterior distribution. In 2010 20th international conference on pattern recognition (pp. 3121-3124). *IEEE*. **doi:10.1109/ICPR.2010.764.**
44. 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.**
45. Rahimi Mashkaleh, M., amirdelavar, M., jamshidi, M., & sharififar, A. (2023). 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]

46. Jensen, J. R. (2005). Introductory image processing: A remote sensing perspective.
47. 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.**
48. 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.**
49. 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.**
50. Li, R., Pan, X., Wu, H., Huang, Y., Li, W., & Li, M. (2021). A comparative study of cast-sensitive methods in digital soil mapping using machine learning algorithms. **doi.org/10.2139/ssrn.4658128 Catena, 208, 105266.**
51. Li, H., Li, J., Zhao, Y., Gong, M., Zhang, Y., & Liu, T. (2021). Cost-sensitive self-paced learning with adaptive regularization for classification of image time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 11713-11727. **doi: 10.1109/ JSTARS.2021.3127754.**
52. Wang, N., Liang, R., Zhao, X., & Gao, Y. (2021). Cost-sensitive hypergraph learning with f-measure optimization. IEEE Transactions on Cybernetics. **doi:10.1109/TCYB.2021.3126756.**
53. Wong, M. L., Seng, K., & Wong, P. K. (2020). Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. *Expert Systems with Applications*, 141, 112918. **doi.org/10.1016/j.eswa. 2019.112918.**
54. Fan, Y., Zhang, C., Liu, Z., Qiu, Z., & He, Y. (2019). Cost-sensitive stacked sparse auto-encoder models to detect striped stem borer infestation on rice based on hyperspectral imaging. *Knowledge-Based Systems*, 168, 49-58. **doi.org/10.1016/j.knosys.2019.01.003.**
55. 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.**