19th Iranian Soil Science Congress
16-18 September, 2025

نوزدهمین کنگره علوم خاک ایران
۱۱ تا ۱۳ آذر ۱۴۰۴

مدیریت جامع‌نگر و هوشمند خاک و آب
Holistic and Smart Soil and Water Management
دانشکدگان کشاورزی و منابع طبیعی دانشگاه تهران
College of Agriculture & Natural Resources, University of Tehran

۰۴۲۵۰-۳۲۰۳۱

افزایش دقت طبقه بندی کلاس های نامتوازن خاک با استفاده از رویکرد یادگیری حساس به هزینه

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

۱-دانش‌آموخته دکتری گروه علوم خاک دانشکده کشاورزی زنجان*(mastanehrahimi@znu.ac.ir)

۲- استاد گروه علوم خاک دانشکده کشاورزی،دانشگاه زنجان

۳- استادیارموسسه تحقیقات خاک وآب،سازمان تحقیقات،آموزش و ترویج کشاورزی،کرج

چکیده

مدیریت بهینه خاک و توسعه پایدار کشاورزی نیازمند دسترسی به اطلاعات دقیق درباره وضعیت و طبقه‌بندی خاک است و پیش‌بینی صحیح کلاس‌های خاک و موقعیت مکانی آن‌ها نقش مهمی در این زمینه دارد. استفاده از روش‌های نوینیادگیری ماشین، به‌ویژهیادگیری حساس به هزینه، می‌تواندبا در نظر گرفتن نامتوازنی توزیع کلاس‌ها، دقت و کارایی پیش‌بینی را بهبود بخشد. در این پژوهش، در اراضی جنوب غربی استان زنجان، ۱۴۸ خاک‌رخ با روش شبکه‌بندی منظم و فاصله متوسط ۵۰۰ متر حفر و تا سطح فامیل طبقه‌بندی شدند. متغیرهای محیطی منتخب شامل داده‌های ژئومورفولوژی، زمین‌شناسی، مدل رقومی ارتفاع و شاخص‌های استخراج‌شده از تصاویر ماهواره‌ای لندست ۸ بودند. مدل‌سازی رابطه خاک و زمین‌نما با الگوریتم جنگل تصادفی و رویکردیادگیری حساس به هزینه در در محیط نرم‌افزار"Rstudio"انجام گرفت.

نتایج نشان داد که خاک‌های منطقه در پنج کلاس نامتعادل شامل تیپیک کلسی‌زرپتز، تیپیک هاپلوزرپتز، جیپسیک هاپلوزرپتز، تیپیک زراورتنتز و لیتیک زراورتنتز قرار دارند. صحت کلی و ضریب کاپا پیش از متعادل‌سازی داده‌ها به‌ترتیب ۶۵ درصد و ۰/۳۲ و پس از متعادل سازیداده‌هابا رویکردیادگیری حساس به هزینهبه‌ترتیب ۸۶ درصد و ۰/۷۷ به دست آمد. مقایسه شاخص‌های صحت کاربر و صحت تولید کننده نشان داد که درحالی‌که جنگل تصادفی در حالت داده‌های نامتعادل در تشخیصدو کلاس اقلیت جیپسیک هاپلوزرپتز و لیتیک زراورتنتز ناکام بود، متعادل سازی داده با رویکرد حساس به هزینه توانست با کاهش خطا، پیش‌بینی دقیقی از این کلاس های اقلیت ارائه دهد.

به طور کلی،یافته‌ها تأیید می‌کند که به‌کارگیرییادگیری حساس به هزینه در کنار الگوریتم جنگل تصادفی به شکل معناداری موجب بهبود پیش‌بینی کلاس‌های خاک، به‌ویژه کلاس‌های با فراوانی کم می‌شود و می‌تواند ابزاری ارزشمند برای تولید نقشه‌های خاک با دقت بالاتر در مدیریت پایدار منابع خاک و کشاورزی باشد.

کلمات کلیدی: جنگل تصادفی، ضریب کاپا، متعادل سازی داده، نقشه برداری رقومی خاک.

مقدمه

خاک به‌عنوان بستر چندمنظوره رشد گیاهان،یکی از ارکان اساسی در دستیابی به اهداف توسعه پایدار محسوب می‌شود (Evans et al., 2022). در شرایط کنونی، بهره‌برداری بهینه از اراضی به‌عنوان یکی از چالش‌های مهم مدیریتی در سطح جهان مطرح است (Garg et al., 2020). یکی از رویکردهای مؤثر در ارتقای دانش خاک و بهبود مدیریت منابع زمین، نقشه‌برداری رقومی کلاس‌های خاک است (Sharififar and Sarmadian, 2023).هرچند نقشه‌های ملی موجود از کلاس‌های خاک کاربردهای ارزشمندی دارند، اما به دلیل بازتاب توزیع نامتعادل داده‌ها در کلاس‌های مختلف و محدودیت در ارائه اطلاعات مکانی دقیق، همواره با چالش‌هایی همراه هستند (رحیمی و همکاران،۱۴۰۲). ازاین‌رو، توسعه نقشه‌های خاک با دقت و وضوح بالاتر ضروری است تا بتوانند مبنای تصمیم‌گیری‌های کارآمد و ارائه دستورالعمل‌های مدیریتی مؤثر برای بهبود مدیریت اراضی قرار گیرند(Helfenstein et al., 2022).

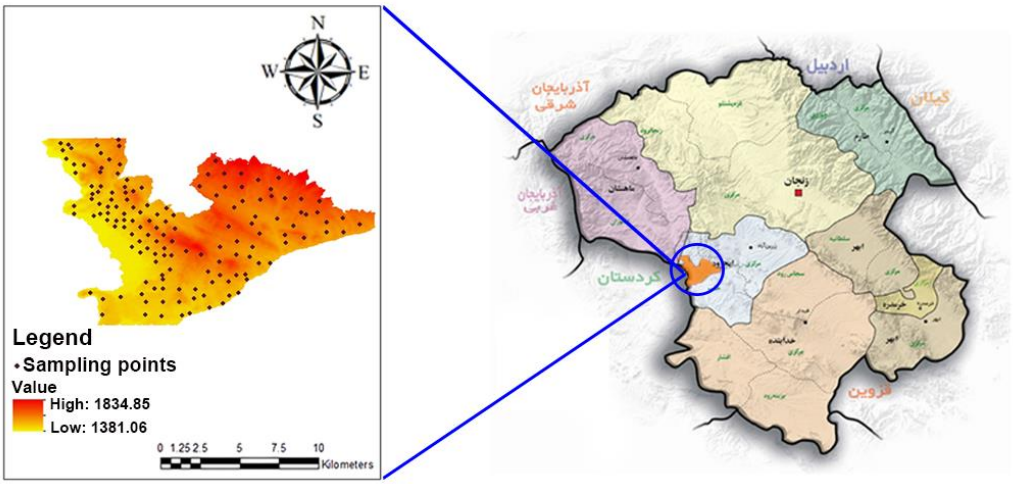
یکی از چالش‌های متداول در مطالعات خاک‌شناسی، مسئله عدم تعادل در تعداد مشاهدات مربوط به کلاس‌های مختلف خاک است. این ناهماهنگی معمولاً تحت تأثیر عواملی شکل می‌گیرد که بر فرآیند تشکیل و تکامل خاک اثرگذارند (Sharififar et al., 2019b ;Taghizadeh-Mehrjardi et al., 2020). به طور کلی، مشکل داده‌های نامتعادل زمانی رخ می‌دهد که یکیا چند کلاس خاک در پهنه جغرافیایی مورد مطالعه به‌طور طبیعی نمونه‌های کمتری نسبت به سایر کلاس‌ها داشته باشند (Fernández et al., 2009). نامتعادل بودن تعداد داده‌ها در کلاس‌های خاک یک منطقه می‌تواند منجر به دست‌یابی کمتر به کلاس‌های اقلیت و برآورد بیش‌ازحد کلاس‌های اکثریت در مدل‌های پیش‌بینی شود. به عبارت دیگر، این پدیده ممکن است باعث شود که بخش‌هایی از منطقه مورد مطالعه، که دارای تعداد مشاهدات خاک‌رخ کمتری هستند، در نقشه‌های رقومی نادیده گرفته شوند (Sharififar et al., 2019a). بیشتر مدل‌هاییادگیری ماشین فرض می‌کنند داده‌ها به‌طور متعادل توزیع شده‌اند؛ بنابراین وقتی با داده‌های نامتعادل آموزش ببینند، نمی‌توانند عملکرد خوبی داشته باشند و نتایج ضعیفی به دست می‌آید (Sharififar and Sarmadian, 2023).

یکی از روش‌های مؤثر برای مقابله با مشکل کلاس‌های نامتعادل در داده‌های خاک، استفاده از رویکردیادگیری حساس به هزینه است. این روش با تمرکز بر کاهش هزینه‌های ناشی از طبقه‌بندی اشتباه، دقت نقشه‌های تولیدشده را افزایش داده و عملکرد بهتری نسبت به توزیع نامتوازن کلاس‌ها ارائه می‌دهد(Vincent et al., 2018 ; Sharififar et al., 2019b).

پژوهش حاضر به بررسی کاربرد این رویکرد در پیش‌بینی مکانی کلاس‌های خاک نامتعادل در بخشی از اراضی جنوب غربی استان زنجان می‌پردازد و نشان می‌دهد که استفاده از یادگیری حساس به هزینه می‌تواند تحلیل داده‌های خاک را دقیق‌تر و کاربردی‌تر کند. این روش نوآورانه، امکان ارائه اطلاعات ارزشمند و قابل اعتماد را برای تصمیم‌گیران و برنامه‌ریزان منطقه فراهم می‌سازد.

مواد و روش‌ها

منطقه مورد مطالعه با مساحت ۱۳۸۲۳ هکتار در غرب استان زنجان و بین مختصات جغرافیایی ۴۷°۲۱′تا۴۸°۱۱′طولشرقی و ۳۱°۳۶′تا۳۶°۳۷′عرضشمالی قرار دارد (شکل ۱). بر اساس داده‌های آماری بلندمدت ۲۰ ساله، متوسط بارندگی سالانه در این منطقه ۳۴۰ میلی‌متر و متوسط دمای سالانه ۱۳ درجه سانتی‌گراد است. ارتفاع متوسط منطقه ۱۴۸۲ متر از سطح دریا بوده و در محدوده ۱۳۷۹ تا ۱۶۹۹ متر متغیر است (سالنامه آماری زنجان، ۱۳۹۸). خاک‌های منطقه دارای رژیم حرارتی مزیک و رژیم رطوبتی زریک هستند و مهم‌ترین واحدهای چشم‌انداز آن تپه‌ماهورها و دشت‌های دامنه‌ای را شامل می‌شوند (موسسه تحقیقات خاک و آب، ۱۳۸۹).پوشش گیاهی منطقه پراکنده و ضعیف بوده و بخش محدودی از اراضی آن به کشاورزی اختصاص دارد. خاک‌های منطقه از رسوبات آبرفتی و آبرفتی-بادرفتی تشکیل شده‌اند که منشأ آن‌ها مارن‌های گچی، سنگ‌آهک و ماسه‌سنگ است.



شکل ۱- موقعیت منطقه موردمطالعه و پراکنش نقاط نمونه‌برداری

در این مطالعه، تعداد ۱۴۸ خاک‌رخ مطابق الگوی طبقه‌بندی تصادفی و با فاصله متوسط حدود ۵۰۰ متر نمونه‌برداری شد (Soil Science Division Staff, 2017). نمونه‌های جمع‌آوری شده پس از خشک شدن در هوا، از الک ۲ میلی‌متری عبور داده شدند و آزمایش‌های فیزیکی و شیمیایی آن‌ها بر اساس روش‌های استاندارد انجام گرفت (USDA, 2004). طبقه‌بندی خاک‌رخ‌ها با استفاده از مشاهدات صحرایی و نتایج آنالیزهای فیزیکی و شیمیایی نمونه‌ها، در چارچوب سیستم جامع رده‌بندی خاک به روش آمریکایی تا سطح فامیل صورت پذیرفت (Soil Survey Staff, 2022).

متغیرهای محیطی شامل داده‌های نقشه‌های ژئومورفولوژی و زمین‌شناسی، تصاویر سنجش از دور و اطلاعات توپوگرافی مورد استفاده قرار گرفتند. نقشه زمین‌شناسی منطقه با مقیاس ۱:۲۵۰۰۰۰، تهیه‌شده توسط سازمان زمین‌شناسی کشور، در محیطArcGISبه فرم رقومی و زمین‌مرجع تبدیل شد (شکل ۲، الف). از مدل رقومی ارتفاع با تفکیک مکانی ۳۰×۳۰ متر (SRTM)، ۱۸شاخص مرتبط با پستی و بلندی‌ها در نرم‌افزارSAGA GISاستخراج گردید. علاوه بر این، ۳۶ شاخص سنجش از دور از تصاویر ماهواره لندست ۸ (OLI/TIRS) با تفکیک مکانی مشابه، پس از اعمال تصحیحات رادیومتریک و اتمسفری در نرم‌افزارENVI، تولید شدند (USGS, 2014). نقشه ژئومورفولوژی منطقه بر اساس ترکیب لایه‌های اطلاعاتی شامل واحدهای لندفرم، مواد مادری و تحلیل تصاویر ماهواره‌ای، با رویکرد سلسله‌مراتبی ارائه شده توسط Zinck (2016) آماده شد.

در ادامه بر اساس رویکرد تحلیل مؤلفه اصلی(PCA) در نرم‌افزار SPSSو رتبه‌بندی اهمیت نسبی مدل یادگیری ماشین از میان ۵۷ متغیر محیطی تولیدشده۱۰ متغیر محیطیشامل اطلاعات نقشه‌های ژئومورفولوژی، اطلاعات زمین‌شناسی و ویژگی‌های مستخرج از مدل رقومی ارتفاع شامل تجزیه‌وتحلیل سایه‌اندازی تپه‌ها[1]، طلوع خورشید[2]، عمق دره[3]، شاخص طول در جهت شیب[4]، فاصله تا شبکه آبراهه[5]، شاخص رطوبتی توپوگرافی[6] و شاخص همواری بالای پشته با درجه تفکیک بالا[7]پس از یکسان‌سازی مقیاس‌ها در محیط نرم‌افزار ArcGISبه‌عنوان مؤثرترین متغیرهای محیطی برای پیش‌بینی کلاس‌های خاک و به‌عنوان ورودی مدل انتخاب گردید (Kuhn and Johnson, 2013).

1- Analytical Hill shading
2- Sunrise
3- Valley Depth
4- LS_Factor
5- Channel Network Distance
6- Topographic Wetness Index (TWI)
7- Multi-Resolution Ridge Top Flatness Index (MRRTF)

مدل جنگل تصادفی یکی از تکنیک‌های پیشرفته یادگیری ماشین و توسعه‌یافته از رویکرد درخت‌های تصمیم برای طبقه‌بندی و رگرسیون است. در این روش، داده‌ها به‌طور مکرر نمونه‌برداری شده و برای شناسایی رابطه بین متغیر وابسته و متغیرهای مستقل و همچنین انجام پیش‌بینی به‌کار گرفته می‌شوند. برخلاف مدل‌های درختی متداول که تنها تعداد محدودی درخت تصمیم ایجاد می‌کنند، در جنگل تصادفی صدها یا حتی هزاران درخت طبقه‌بندی ساخته می‌شود (Breiman and Cutler, 2004).

در رویکرد حساس به هزینه، هدف الگوریتم به حداقل رساندن هزینه ناشی از خطاهای طبقه‌بندی است. یکی از کاربردهای این رویکرد در نقشه‌برداری رقومی خاک، استفاده از مدل‌های وزن‌دار جنگل تصادفی است. جنگل تصادفی به‌عنوان یک الگوریتم یادگیری گروهی، با ساخت تعداد زیادی درخت تصمیم در فرایند آموزش عمل می‌کند (Zhao et al., 2018). برای حساس‌سازی آن نسبت به هزینه، وزن‌هایی متناسب با معکوس توزیع کلاس‌ها به نمونه‌ها اختصاص داده می‌شود تا توجه مدل بیشتر بر روی کلاس‌های اقلیت متمرکز شود (Zhang et al., 2021). در این مطالعه، پیاده‌سازی مدل با استفاده از تابع rfدر بسته Random Forest در محیط نرم افزار RStudio انجام شد.

به منظور ارزیابی صحت مدل مورداستفاده، داده‌ها به‌طور تصادفیبه دو دسته داده‌های آموزشی و اعتبارسنجی تقسیمشدند. ۸۰ درصد داده‌ها برای آموزش مدل و ۲۰ درصد دیگر به‌عنوان داده‌های اعتبارسنجی برای ارزیابی مورداستفاده قرار گرفتند (رحیمی و همکاران، ۱۴۰۲). هر مدل با داده‌های آموزشی برازش داده شد و سپس پیش‌بینیبرایداده‌های اعتبارسنجی انجام شد. کلاس‌های پیش‌بینی‌شدهبااستفادهازماتریسخطابرحسبدرصدبیان شد. پارامترهای استخراج‌شده از ماتریسخطاشامل صحت کلی نقشه (Overall accuracy) ، صحت تولیدکننده (Producer accuracy) ، صحت کاربر (Users accuracy) و ضریبکاپا (Kappa index)برای اعتبارسنجی مورداستفاده قرار گرفت (Congalton, 1991 ; Jensen, 2005).

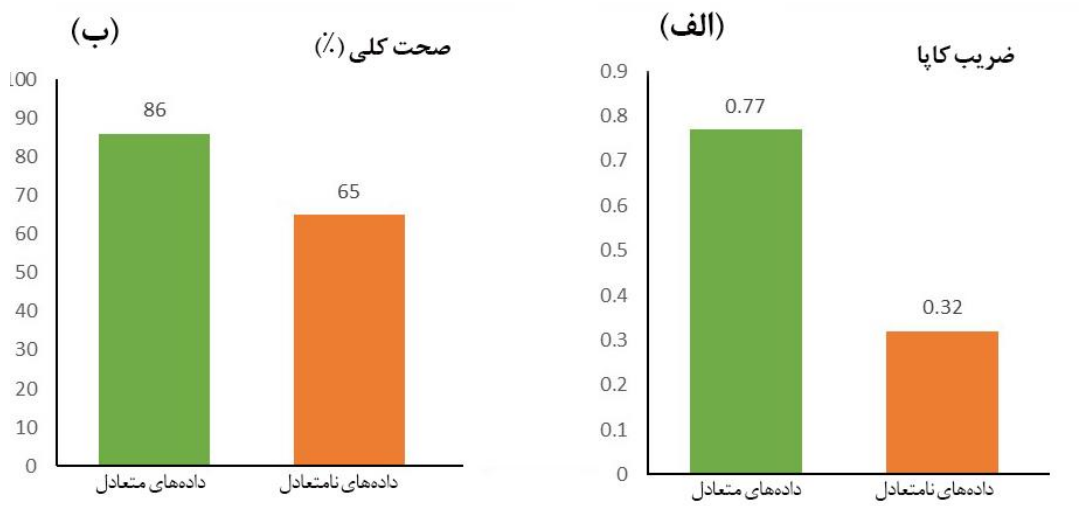
نتایج و بحث

نتایج طبقه‌بندی نشان داد که خاک‌های منطقه در دو رده انتی‌سولز و این‌سپتی‌سولز قرار می‌گیرند. این خاک‌ها در سطح زیرگروه به پنج کلاس تقسیم شدند: تیپیک کلسی‌زرپتز(A) ، تیپیک هاپلوزرپتز(B) ، جیپسیک هاپلوزرپتز(C) ، تیپیک زراورتنتز (D) و لیتیک زراورتنتز (E) . جدول (۱) تعداد نمونه‌های هر کلاس را پیش و پس از متعادل‌سازی داده‌ها نشان می‌دهد. بر اساس فراوانی، زیرگروه‌هایA ، B وD به‌عنوان کلاس‌های اکثریت و زیرگروه‌های C و E به‌عنوان کلاس‌های اقلیت شناسایی شدند. پس از اعمال متعادل‌سازی، تعداد نمونه‌های کلاس‌های اکثریت کاهش یافت و در مقابل، نمونه‌های کلاس‌های اقلیت حدود دو تا سه برابر افزایش یافتند.

جدول۱- توزیع فراوانی کلاس‌های خاک قبل و بعد از متعادل‌سازی

کد کلاس	زیرگروه‌های خاک	تعداد مشاهده‌ها قبل از متعادل‌سازی	تعداد مشاهده‌ها بعد از متعادل‌سازی
A	تیپیک کلسی‌زرپتز	۶۸	۴۷
B	تیپیک هاپلوزرپتز	۲۶	۲۰
C	جیپسیک هاپلوزرپتز	۱۲	۲۵
D	تیپیک زراورتنتز	۳۱	۲۳
E	لیتیک زراورتنتز	۱۱	۳۳

نتایج صحت‌سنجی پیش‌بینی مکانی کلاس‌های خاک در شرایط معمول و پس از رفع عدم‌تعادل داده‌ها با استفاده از الگوریتم جنگل تصادفی و رویکرد یادگیری حساس به هزینه در شکل (۲) ارائه شده است. پیش از متعادل‌سازی داده‌ها، شاخص کاپا و صحت کلی به‌ترتیب ۰/۳۲ و ۶۵درصد بود، در حالی که پس از اعمال رویکرد حساس به هزینه این مقادیر به ۰/۷۷و ۸۶درصد افزایش یافت. به این ترتیب، صحت کلی حدود ۲۱ درصد و ضریب کاپا بیش از دو برابر بهبود پیدا کرد.



شکل۲_ دقت پیش‌بینی طبقه‌بندی زیرگروه‌های خاک قبل و بعد از متعادل‌سازی داده‌ها

این نتایج نشان می‌دهد که متعادل‌سازی داده‌های نامتوازن با رویکرد یادگیری حساس به هزینه، دقت پیش‌بینی کلاس‌های خاک و کیفیت نقشه تولیدشده را به‌طور چشمگیری ارتقا می‌دهد. در واقع، تمرکز مدل بر روی کلاس‌های اقلیت باعث کاهش خطای پیش‌بینی و افزایش دقت کلی مدل شده است.

Mienye و Sun (2021) در پژوهش خود کارایی روش‌های یادگیری حساس به هزینه را در مواجهه با داده‌های نامتعادل پزشکی بررسی کردند. آن‌ها چهار الگوریتم شامل رگرسیون لجستیک حساس به هزینه، درخت تصمیم حساس به هزینه، گرادیان تقویتی تصادفی حساس به هزینه و جنگل تصادفی حساس به هزینه را مقایسه نمودند و نشان دادند که الگوریتم جنگل تصادفی حساس به هزینه، دقت و کارایی به‌مراتب بالاتری نسبت به سایر روش‌ها در بهبود داده‌های نامتعادل دارد.

همچنین Kang و همکاران (2022) و Devi و همکاران (2019) در مطالعات مستقل خود استفاده از جنگل تصادفی حساس به هزینه را بررسی کرده و گزارش دادند که این روش می‌تواند دقت صحت‌سنجی نتایج را تا حدود ۳۰درصد افزایش دهد.یافته‌های آن‌ها بیانگر آن است که این الگوریتم نه‌تنها در مقایسه با جنگل تصادفی استاندارد عملکرد بهتری دارد، بلکه برای مجموعه داده‌های کوچک نیز کارآمد و قابل‌استفاده است.

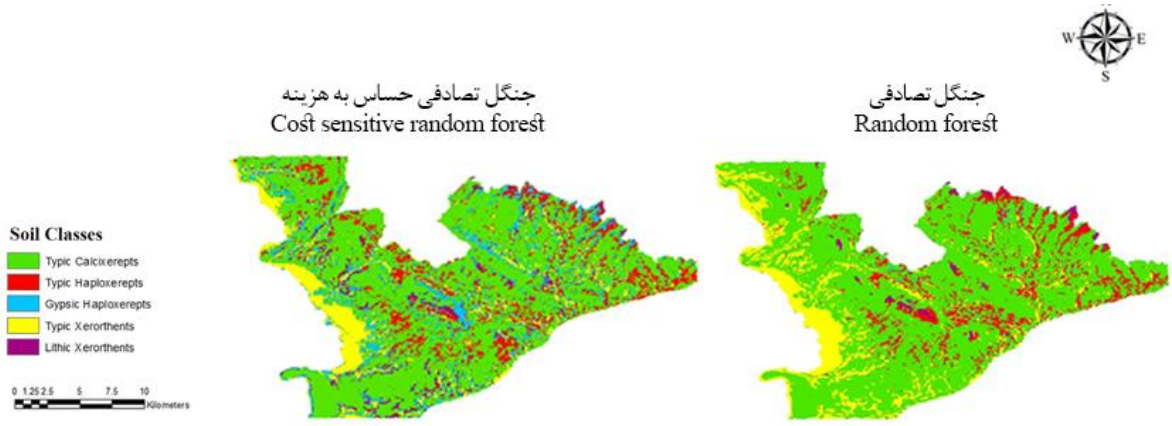
جدول (۲) نتایج صحت کاربر و صحت تولیدکننده را برای کلاس‌های خاک در شرایط معمول و پس از به‌کارگیری رویکرد یادگیری حساس به هزینه با الگوریتم جنگل تصادفی نشان می‌دهد. اعتبارسنجی نتایج بیانگر آن است که این رویکرد دقت پیش‌بینی تمامی زیرگروه‌های خاک را افزایش داده است. به‌ویژه دو کلاس کم‌رخداد جیپسیک هاپلوزرپتز و لیتیک زراورتنز که در حالت نامتعادل توسط مدل شناسایی نشده بودند، پس از اعمال یادگیری حساس به هزینه با دقت مناسبی پیش‌بینی شدند.

جدول۲- صحت تولیدکننده و کاربر در سطح زیرگروه‌های خاک، قبل و بعد از متعادل‌سازی داده‌ها

صحت تولیدکننده (٪)		**صحت کاربر (٪)**		**قابلیتاطمینان**
داده های متعادل	داده‌های نامتعادل	داده های متعادل	داده‌های نامتعادل	مدل‌های یادگیریماشین
۱۰۰	۸۵	۹۵	۶۱	تیپیک کلسی زرپتز
۸۳	۵۰	۷۱	۱۰۰	تیپیک هاپلوزرپتز
۸۵	۰	۱۰۰	NaN	جیپسیک هاپلوزرپتز
۱۰۰	۳۴	۸۱	۶۵	تیپیک زراورتنتز
۹۱	۰	۱۰۰	NaN	لیتیک زراورتنتز

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

همانطور که در شکل (۳) مشاهده می‌شود، اعمال یادگیری حساس به هزینه همراه با متعادلسازی داده‌ها در الگوریتم جنگل تصادفی، منجر به بهبود چشمگیر در پیش‌بینی کلاس‌های اقلیت (جیپسیک هاپلوزرپتز و لیتیک زراورتنتز) شده است. این در حالی است که مدل آموزش‌دیده با داده‌های نامتعادل، قادر به شناسایی این دو کلاس نبود.مطالعات Fernández و همکاران (2009) نشان می‌دهند که روش‌های حساس به هزینه ابزار مؤثری برای رفع مشکل عدم تعادل در داده‌کاوی و یادگیری ماشین هستند. بر اساس مطالعات محققین دیگر، این رویکرد در نقشه‌برداری رقومی خاک اهمیت ویژه‌ای دارد، زیرا تضمین می‌کند تمامی کلاس‌های خاک، صرف‌نظر از فراوانی آن‌ها، در نقشه‌ها به‌طور دقیق نمایش داده شوند. در نتیجه، دقت و قابلیت اطمینان مدل‌های نقشه‌برداری خاک ارتقا یافته و امکان تصمیم‌گیری و مدیریت زمین به شکل کارآمدتری فراهم می‌شود(Zhang et al., 2018؛ Wong et al, 2020 و Li et al., 2021).



شکل۳- نقشه‌های حاصل از الگوریتم‌های یادگیری ماشین قبل و پس از اعمال متعادل‌سازی داده‌ها با رویکرد یادگیری حساس به هزینه

نتیجه‌گیری

استفاده از الگوریتم‌های متداول در نقشه‌برداری رقومی خاک، بدون در نظر گرفتن نامتعادل بودن فراوانی کلاس‌ها، معمولاً باعث برآورد بیش‌ازحد کلاس‌های پرتکرار و در مقابل نادیده گرفتن یا حذف کلاس‌های کم‌رخداد می‌شود. یافته‌های این پژوهش نشان داد که توزیع نامتعادل کلاس‌ها اثر مستقیم بر عملکرد مدل‌های پیش‌بینی دارد. به‌کارگیری رویکرد یادگیری حساس به هزینه در الگوریتم جنگل تصادفی موجب شد دقت مدل در تشخیص کلاس‌های اقلیت به‌طور معناداری افزایش یابد و کیفیت نقشه‌های تولیدی بهبود پیدا کند.این نتایج بیانگر آن است که رفع مشکل عدم تعادل کلاس‌ها گامی کلیدی برای ارتقای کارایی مدل‌های طبقه‌بندی خاک است. با توجه به محدود بودن مطالعات انجام‌شده در این زمینه، نتایج این تحقیق می‌تواند به‌عنوان الگویی ارزشمند برای پژوهش‌های آتی در نقشه‌برداری رقومی خاک مورد استفاده قرار گیرد و به بهبود دقت مدل‌ها و تفسیر واقع‌بینانه‌تر واحدهای خاک در مقیاس منطقه‌ای کمک کند.

فهرست منابع

رحیمی مشکله، م.، دلاور، م.ا.، جمشیدی،م.،شریفی فر، ا. (۱۴۰۲). بهبود طبقه بندی داده های نامتعادل خاک با استفاده از الگوریتم های یادگیری ماشین در بخشی از اراضی استان زنجان.مهندسیزراعی. ۴۶(۱): ۶۱-۸۲.

سالنامه آماری استان زنجان.(۱۳۹۸). «۱- سرزمین و آب‌وهوا»، سازمان آمار کشور.

موسسه تحقیقات خاک و آب. (۱۳۸۹). مطالعات پژوهشی مکانیابی، خاکشناسی و ارزیابی اراضی برای احداث باغات در استان زنجان. نشریه شماره ۱۵۴۷، ۳۶۴ صفحه. کرج. ایران.

Breiman, L., & Cutler, A. (2004). Random Forests. Department of Statistics, University of Berkeley.

Congalton, R.G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. Remote sensing of environment, 37(1): 35-46.

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

Evans, D.L., Janes-Bassett, V., Borrelli, P., Chenu, C., Ferreira, C.S., Griffiths, R.I., Kalantari, Z., Keesstra, S., Lal, R., Panagos, P. and Robinson, D.A. (2022). Sustainable futures over the next decade are rooted in soil science. European Journal of Soil Science, 73(1): e13145.

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.

Garg, K.K., Anantha, K.H., Nune, R., Akuraju, V.R., Singh, P., Gumma, M.K., Dixit, S., 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.

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.

Jensen, J.R. (1996). Introductory digital image processing: a remote sensing perspective (No. Ed. 2). Prentice-Hall Inc...

Kuhn, M., Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

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.

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.

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.

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.

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.

Sharififar, A., Sarmadian, F. (2023). Coping with imbalanced data problem in digital mapping of soil classes. European Journal of Soil Science, 74(3): e13368.

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.

Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., Toomanian, N., Scholten, T. (2020). Synthetic resampling strategies and machine learning for digital soil mapping in Iran. European Journal of Soil Science, 71(3): 352-368.

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.

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.

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.

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.

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.

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.

Improving Classification Accuracy of Imbalanced Soil Classes Using a Cost-Sensitive Learning Approach

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

1- Ph.D. Graduate, Soil Science Department, Faculty of Agriculture, University of Zanjan, Iran * (mastanehrahimi@znu.ac.ir)

2- Professor, Soil Science Department, Faculty of Agriculture, University of Zanjan, Iran

3 -Assistant Professor, Soil and Water Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran

Abstract

Optimal soil management and sustainable agricultural development require accurate information on soil status and classification, while precise prediction of soil classes and their spatial distribution plays a crucial role in this regard. The use of advanced machine learning methods, particularly cost-sensitive learning, can improve prediction accuracy and efficiency by accounting for imbalanced class distributions. In this study, 148 soil profiles were systematically sampled at 500-meter intervals in the southwestern region of Zanjan province and classified up to the family level. Selected environmental variables included geomorphological and geological data, digital elevation models (DEM), and indices extracted from Landsat 8 imagery. Soil-landscape modeling was performed using the Random Forest algorithm (RF) coupled with a cost-sensitive learning approach in the RStudio environment.

Results indicated that soils in the study area belonged to five imbalanced classes: Typic Calcixerepts, Typic Haploxerepts, Gypsic Haploxerepts, Typic Xerorthents, and Lithic Xerorthents. Overall accuracy (OA) and the Kappa coefficient before data balancing were 65% and 0.32, respectively, whereas after data balancing using the cost-sensitive approach, they improved to 86% and 0.77. Comparison of user and producer accuracies (UA, PA) revealed that, while Random Forest failed to accurately predict the minority classes Gypsic Haploxererts and Lithic Xerorthents under imbalanced data conditions, the cost-sensitive balancing approach substantially reduced errors and provided reliable predictions for these minority classes.

Overall, the findings demonstrate that applying cost-sensitive learning alongside the Random Forest algorithm significantly enhances soil class prediction, particularly for low-frequency classes, and can serve as a valuable tool for producing higher-accuracy soil maps for sustainable soil and agricultural resource management.

Keywords: Random Forest, Kappa coefficient, Data balancing, Digital soil mapping.