Research Article

Agricultural Engineering., 46(1) (2023) 61-81 ISSN (P): 2588-526X

DOI: 10.22055/AGEN.2023.43838.1667 ISSN (E): 2588-5944

Improving the classification of Soil imbalanced data using machine learning algorithms

M. Rahimi Mashkaleh[1], M.A. Delavar[2], M.Jamshidi[3] and A. Sharififar[4]

1. Ph.D. Student of Department of Soil Science, Faculty of Agriculture, University of Zanjan, Iran.
2. Associate Professor, Department of Soil Science, Faculty of Agriculture, University of Zanjan, Iran.
3. Assistant Professor, Soil and Water Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran.
4. Researcher, Department of Soil Science, Faculty of Agriculture, University of Tehran, Iran.

Received: 20 April 2023 Accepted: 18 June 2023

Abstract

Introduction: Digital soil maps are commonly used in soil science studies, but the imbalance in soil class distribution can limit the performance of machine learning algorithms. This has led to increased attention from researchers on solutions to overcome these limitations. This study aims to improve the classification of imbalanced soil data using the resampling pre-treatment technique in three prediction models: Random Forest (RF), Boosted Regression Trees (BRT), and Multinomial Logistic Regression (MNLR), in a region of Zanjan province in Iran.

Materials and Methods: Soil sampling was done based on a regular grid pattern with 500-meter average intervals, and 148 soil profiles samples were randomly selected and classified into five classes at the subgroup level. Environmental covariates, including geomorphological and geological maps, a digital elevation model (DEM), and remote sensing (RS) were selected using principal component analysis (PCA) and expert knowledge methods. The most effective environmental variables for predicting soil classes were selected and used as input to the models. Extraction of environmental covariates was done using ENVI and SAGA_GIS software, and modeling of soil-landscape relationships was done using the aforementioned algorithms in Rstudio software. The resampling technique was applied to the minority and majority soil classes prior to modeling.

Results and Discussion: The use of imbalanced data for mapping resulted in a loss of minority classes and relatively low Kappa agreement values and overall accuracy for RF (overall=65%, k=0.32) and BRT models (overall=60%, k=0.35), indicating decreased accuracy in spatial predictions of soil subgroups. However, after resampling the data, the overall accuracy and Kappa coefficient statistics increased in all models, resulting in improved spatial predictions. The BRT model provided an acceptable estimate by maintaining the minority classes and had a Kappa coefficient of 0.64 and an overall accuracy of 75% in the spatial prediction of soil subgroups. The producer accuracy (PA) and user accuracy (UA) results showed that the two classes of Gypsic Haploxerepts and Lithic Xerorthents, which were excluded when training using imbalanced datasets in RF and BRT algorithms, showed significant improvement after balancing the data. Results showed that they were well-predicted in the RF algorithm (UA = 100%, 78%) and BRT algorithm (UA= 60% and 70%) using treated data, highlighting the importance of balancing the

data for improved spatial predictions. Additionally, these minority classes showed producer accuracy in the RF algorithm (PA= 75%, 88%) and BRT algorithm (PA= 100%, 78%) in comparison to zero accuracy when training using imbalanced data. On the other hand, the validation results of the MNLR algorithm showed that despite maintaining the minority classes after balancing the data, the minority classes were predicted with less accuracy, indicating the need for further improvement in spatial predictions using this algorithm.

Conclusion: The results of this study demonstrate that using imbalanced distribution of class observations in modeling can lead to uncertain digital soil maps with lost minority classes and relatively poor accuracies. Therefore, it is crucial for researchers to address this issue by applying data resampling techniques with over- and under-sampling to balance the class distribution in the data before modeling. In addition to resampling techniques, there are other ways to address imbalanced class distribution, such as cost-sensitive learning, ensemble learning, and threshold-moving. Researchers can explore these methods to further improve the accuracy of soil classification models.

Keywords: Boosted regression trees, data pretreatment, oversampling, resampling methods, minority class

مهندسی زراعی (مجله علمی کشاورزی) جلد ۴۶ شماره ۱، بهار ۱۴۰۲

مقاله پژوهشی

بهبود طبقه‌بندی داده‌های نامتعادل خاک با استفاده از الگوریتم‌های یادگیری ماشین در بخشی از اراضی استان زنجان

مستانه رحیمی مشکله[۱]، محمد امیر دلاور[۲*]، محمد جمشیدی[۳] و امین شریفی‌فر[۴]

۱- دانشجوی دکتری گروه علوم خاک دانشکده کشاورزی، دانشگاه زنجان، ایران
۲- دانشیار گروه علوم خاک دانشکده کشاورزی، دانشگاه زنجان، ایران
۳- استادیار موسسه تحقیقات خاک و آب، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران
۴- پژوهشگر گروه علوم خاک پردیس کشاورزی، دانشگاه تهران، ایران

تاریخچه مقاله

دریافت: ۱۴۰۲/۰۱/۳۱

پذیرش نهایی: ۱۴۰۲/۰۳/۲۸

کلمات کلیدی:
واژه‌های کلیدی:
بیش نمونه‌گیری،
پیش تیمار داده،
درخت تصمیم توسعه یافته،
کلاس اقلیت،
نمونه‌برداری مجدد

* عهده دار مکاتبات
E-mail: amir-delavar@znu.ac.ir:

چکیده

علی‌رغم استفاده گسترده از روش‌های نقشه‌برداری رقومی خاک در مطالعات خاکشناسی، محدودیت‌های مربوط به عدم تعادل در دامنه گسترش کلاس‌های خاک مانع عملکرد موفقیت‌آمیز بسیاری از الگوریتم‌های یادگیری ماشین در این روش‌ها شده است و اخیراً توجه محققان زیادی به ارائه راهکارهایی برای رفع این محدودیت‌ها جلب شده است. هدف از انجام این پژوهش بهبود طبقه‌بندی داده‌های نامتعادل خاک با استفاده از روش پیش‌درمانی نمونه‌گیری مجدد با استفاده از سه مدل پیش‌بینی جنگل تصادفی، درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای در بخشی از اراضی استان زنجان است. برای این منظور موقعیت ۱۴۸ خاک‌رخ مشاهداتی بر اساس الگوی شبکه‌بندی منظم با فاصله ۵۰۰ متر حفر و پس از انتقال به آزمایشگاه تجزیه‌های فیزیکی و شیمیایی شامل بافت خاک، پ‌هاش خاک، کربنات کلسیم معادل، ظرفیت تبادل کاتیونی، قابلیت هدایت الکتریکی، کربن آلی و گچ بر روی تمام نمونه‌ها انجام و بر اساس استانداردهای سیستم جامع رده‌بندی خاک تشریح و طبقه‌بندی گردید. متغیرهای محیطی شامل اطلاعات نقشه‌های ژئومورفولوژی و زمین-شناسی، 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 گردید. مدل‌سازی رابطه خاک - زمین‌نما با استفاده از الگوریتم‌های جنگل تصادفی، درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای در محیط

نرم‌افزار "Rstudio" انجام شد. خاک‌های منطقه در سطح زیرگروه در پنج کلاس با توزیع نامتعادل شامل تیپیک کلسی‌زرپتز، تیپیک هاپلوزرپتز، جیپسیک هاپلوزرپتز، تیپیک زراورتنتز و لیتیک زراورتنتز قرار گرفتند. بر اساس یافته‌های این پژوهش مدل‌سازی با استفاده از داده‌های نامتعادل، منجر به از دست دادن کلاس‌های با مشاهده‌های کم تعداد و کاهش مقادیر صحت کلی برای مدل‌های جنگل تصادفی، درخت تصمیم و مدل رگرسیون لجستیک چند جمله‌ای به ترتیب ٦٥، ٦٠ و ٦٦ درصد و شاخص کاپا به ترتیب ۰/۳۲، ۰/۳۵ و ۰/٤١ بود. در شرایط داده‌های نامتعادل مدل رگرسیون لجستیک چندجمله‌ای بالاترین دقت را نسبت به دو مدل دیگر نشان داد. پس از نمونه‌برداری مجدد داده‌ها در قالب فرآیند متعادل‌سازی، مقادیر صحت کلی در سه مدل جنگل تصادفی، درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای به ترتیب ۷۱، ۷۵ و ۷۰ درصد و مقادیر شاخص کاپا ۰/٥٤، ۰/٦٤ و ۰/٤٧ بود که در هر سه مدل افزایش پیدا کرد. در شرایط متعادل‌سازی داده‌ها مدل درخت تصمیم توسعه‌یافته با حفظ کلاس‌های کم تعداد در پیش‌بینی مکانی زیرگروه‌های خاک، برآورد قابل قبولی ارائه داد. نتایج ارزیابی مدل‌ها نشان داد مدل‌سازی با استفاده از توزیع نامتعادل کلاس‌های خاک به دلیل از دست رفتن کلاس‌های با تعداد مشاهده کمتر منجر به ایجاد نقشه‌های نامطمئن و با دقت نسبتاً ضعیف می‌شود، متعادل‌سازی داده‌ها، با استفاده از روش نمونه‌برداری مجدد می‌تواند منجر به بهبود قابل‌توجهی در حفظ کلاس‌های خاک با تعداد مشاهده کمتر و ارتقاء دقت مدل‌های مبتنی بر روابط خاک - زمین‌نما در مطالعات نقشه‌برداری رقومی خاک گردد.

مقدمه

تولید نقشه‌های دقیق خاک برای مدیریت صحیح کشاورزی و حفاظت از محیط‌زیست ضروری است (۲۷). نقشه-های خاک که نقش مؤثر و مهمی در تعیین راهکارها و برنامه-های مناسب مدیریت زمین مبتنی بر شرایط و قابلیت‌های خاص انواع خاک‌ها دارند (۳۶). نقشه‌های خاک می‌توانند پیش‌بینی رفتار خاک‌ها در کاربری‌های مختلف را فراهم کرده و قابلیت‌های کاربردی خاک را در اختیار کاربران قرار می‌دهد (۴). روش‌های نقشه‌برداری رقومی خاک پتانسیل زیادی برای غلبه بر برخی از محدودیت‌های نقشه‌برداری مرسوم خاک نشان داده است (۸ و ۲۳). نقشه‌برداری رقومی خاک همراه با اطلاعات محیطی و استنباط از طریق مدل‌های خاک، می‌تواند برای ارزیابی خصوصیات کلاس‌های خاک استفاده شود (۱۱). در دهه‌های گذشته، جامعه پویا و رو به رشد، روش‌های نقشه‌برداری رقومی خاک را توسعه و به اشتراک گذاشته شده است تا با کمک نقشه‌برداری رقومی خاک وارد عصر جدیدی از عملیاتی شدن شود (۲۵)؛ 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 (۲۷).

امروزه با توجه به گسترش روزافزون اطلاعات دقیق در مورد منابع سرزمینی، بهره‌گیری از روش‌هایی همچون داده‌کاوی برای استخراج دانش و اطلاعات نهفته در داده‌ها، امری غیرقابل‌اجتناب است؛ اما با توجه به ماهیت توزیع خاک‌ها، مسئله مهم در مدل‌سازی و در فرآیند نقشه‌برداری کلاس‌های خاک، عدم تعادل کلاس‌های مشاهده‌شده است (۳۵). داده‌های نامتعادل[1] زمانی بروز می‌کنند که نمونه‌های یک یا چند کلاس به‌طورطبیعی در پهنه جغرافیایی موردمطالعه تعداد کمتری از سایر کلاس‌ها داشته باشد (۱۳). استفاده از داده‌های آموزشی نامتعادل ممکن است باعث برآورد دقت غیرقابل‌اعتماد شده و کلاس‌هایی که در منطقه مطالعاتی دارای تعداد محدود هستند اغلب طبقه‌بندی آن‌ها با پیش‌بینی نادرست همراه و یا نادیده گرفته شوند. این مسئله به‌عنوان یک مشکل عدم تعادل کلاس[2] در یادگیری ماشین شناخته می‌شود و مجموعه داده‌هایی که این معیار را برآورده نمی‌کنند به‌عنوان داده‌های نامتعادل نامیده می‌شوند. بیشتر مجموعه داده‌های کلاس‌های خاک، داده‌های نامتعادل هستند که این مسئله منجر به کاهش دقت کلی نقشه خاک شده و در بسیاری از مواقع حذف واحد اقلیت را به همراه دارد (۴۰). الگوریتم‌های یادگیری ماشین معمول، اغلب میزان صحت بالایی برای داده‌های اکثریت[3] به دست می‌آورند درصورتی‌که برای داده‌های اقلیت[4]، عکس آن است (۳۱). تعداد داده‌های نامتعادل در کلاس‌های مشاهده‌شده خاک یک منطقه می‌تواند منجر به برآورد کم‌تر از حد کلاس‌های اکثریت در مدل‌سازی و تخمین بیش‌ازحد طبقات اکثریت در مدل‌سازی پیش‌گویانه (تخمینی) شود. به‌عبارت‌دیگر، اثر این پدیده در نهایت به‌گونه‌ای پیش می‌رود که یک محدوده از منطقه موردمطالعه با مشاهده‌های خاک‌رخ کمتر ممکن است در نقشه‌های رقومی حذف گردد (۳۵). همچنان که در مدل‌سازی گروه‌های خاک مرجع ایران نشان داده شد که گروه‌های خاک مرجع با تعداد نمونه بیشتر با الگوریتم‌های یادگیری ماشین با دقت بیشتری نسبت به کلاس‌هایی با تعداد نمونه کمتر مدل‌سازی می‌شوند (۴۲). رویکردهای متفاوتی که برای غلبه بر مشکل عدم توازن کلاس‌های خاک وجود دارد، اما تحقیقات معدودی در این خصوص انجام شده است. با مرور یافته‌های تحقیقاتی مشخص شد که این چالش هم در الگوریتم‌های یادگیری ماشین (۳۵ و ۳۶) و هم در مدل‌های شبیه‌سازی مکانی (۴۰) وجود دارد. روش نمونه‌گیری مجدد[5] می‌تواند عدم تعادل داده‌های آموزشی با تعدیل تعداد نمونه‌ها از کلاس‌های اکثریت و اقلیت، داده‌های آموزشی را متعادل کند (۱۹). نمونه‌برداری مجدد با روش‌های نمونه‌زدایی از کلاس اکثریت (کم‌نمونه‌گیری[6])، نمونه‌افزایی به کلاس اقلیت (بیش‌نمونه‌گیری[7]) و یا ترکیبی از هر دو روش انجام می‌شود (۵). کم‌نمونه‌گیری می‌تواند با حذف برخی از موارد اکثریت از داده‌های آموزشی، موارد اکثریت و اقلیت را متعادل کند. از طرف دیگر، بیش‌نمونه‌گیری می‌تواند با ضمیمه کردن برخی از مواردی که ویژگی‌های آن‌ها مانند اقلیت در داده‌های آموزشی است، به همان شیوه کم‌نمونه‌برداری عمل کند (۳۳). تقی‌زاده مهرجردی و همکاران[8] (۴۱) برای حل مشکل داده‌های نامتعادل از هشت روش نمونه‌گیری مجدد (نمونه‌برداری بیش‌ازحد و کمتر از حد تصادفی، نمونه‌برداری بیش‌ازحد از اقلیت مصنوعی، نمونه‌گیری مصنوعی تطبیقی، نویز گوسی[9]، تومک، نزدیک‌ترین همسایگان متراکم و روش انتخاب یک‌طرفه) استفاده کردند و نتیجه گرفتند متعادل‌سازی داده‌ها با استفاده از الگوریتم‌های نمونه‌گیری مصنوعی می‌تواند باعث افزایش دقت پیش‌بینی کلاس‌های خاک شده و به‌طور قابل‌توجهی دقت پیش‌بینی در نقشه‌برداری رقومی خاک را بهبود بخشد. هانگ و همکاران (۱۵) یک روش نمونه‌گیری بیش‌ازحد تصادفی را برای تولید داده‌های متعادل خاک با استفاده از الگوریتم‌های مختلف یادگیری ماشین موردمطالعه قرار دادند. آن‌ها دریافتند استفاده از روش نمونه‌گیری بیش‌ازحد تصادفی موجب افزایش دقت تولید نقشه برای گروه بزرگ‌های

1- Imbalanced data
2- Class imbalance
3- Majority class
4- Minority class

5- Resampling data
6- Undersampling
7- Oversampling
8- Taghizadeh-Mehrjardi *et al.*
9- Gaussian noise

خاک با استفاده از الگوریتم‌های مختلف یادگیری ماشین می‌گردد. شریفی‌فر و همکاران[1] (۳۶) دو روش نمونه‌برداری مجدد نمونه‌برداری بیش‌ازحد تصادفی و نمونه‌برداری کمتر از حد تصادفی را برای مقابله با مسئله داده‌های نامتعادل خاک مورد ارزیابی قرار دادند. آن‌ها نشان دادند که متعادل‌سازی کلاس‌های خاک منجر به کاهش قابل‌توجه عدم قطعیت الگوریتم‌های یادگیری ماشین می‌شود.

بااینکه مطالعات در زمینه داده‌های نامتعادل، در برخی علوم از جمله رایانه به‌وفور انجام‌شده اما مطالعات اندکی در رابطه با روش‌های مختلف متعادل‌سازی داده‌ها برای علوم خاک وجود دارد که به‌طور عمده در مقیاس کوچک و تعداد محدودی از روش‌های متعادل‌سازی داده‌ها هستند. در واقع مطالعات خاکشناسی با استفاده از داده‌های نامتعادل در ایران بسیار کم انجام شده است؛ بنابراین مطالعه حاضر سعی دارد ضمن استفاده از داده‌های میدانی و اطلاعات نقشه‌های خاک موجود و بر اساس روابط خاک – زمین‌نما و مقایسه روش‌های یادگیری ماشین برای ارتقای سطح دقت این نقشه‌ها، به بررسی چگونگی حل مشکل داده‌های نامتعادل خاک با استفاده از روش نمونه-گیری مجدد در بخشی از اراضی استان زنجان بپردازد.

مواد و روش‌ها

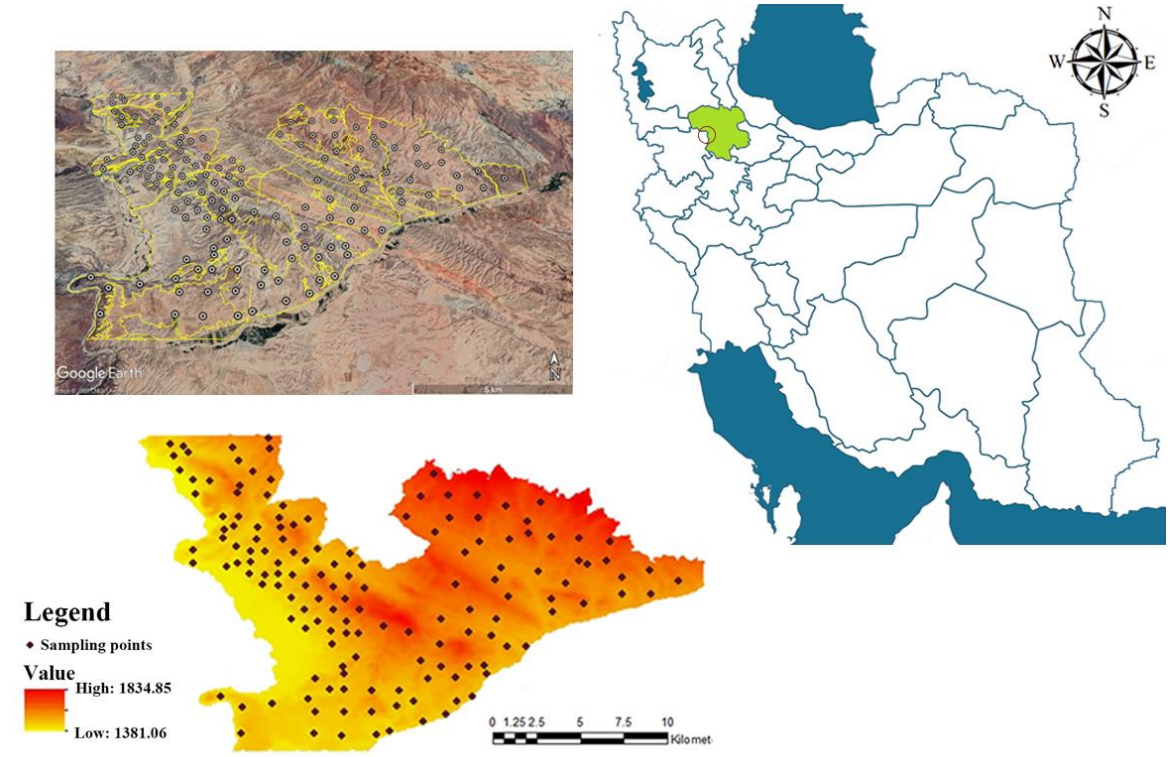
ویژگی‌های منطقه موردمطالعه

منطقه موردمطالعه به مساحت ۱۳۸۲۳ هکتار در جنوب غربی استان زنجان در مختصات جغرافیائی ۴۷ درجه و ۹۱ دقیقه تا ۴۸ درجه و ۱۱ دقیقه طول شرقی و ۳۶ درجه و ۳۷ دقیقه تا ۳۶ درجه و ۳۱ دقیقه عرض شمالی واقع شده است (شکل ۱). با توجه به آمار بلندمدت ۲۰ ساله، متوسط بارندگی سالانه ۳۴۰ میلی‌متر و متوسط دمای سالانه ۱۴ درجه سلسیوس است. میانگین ارتفاع منطقه ۱۴۸۲ متر از سطح دریا بین ۱۳۷۹ تا ۱۶۹۹ متر متغیر است (۳۹). خاک‌های منطقه دارای رژیم حرارتی مزیک[2] و رژیم رطوبتی زریک[3] 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[4] و دشت‌های دامنه‌ای[5] است (۳۷). منطقه موردمطالعه دارای پوشش گیاهی تنک بوده و در زمره مراتع ضعیف قرار می‌گیرند و بخش کمی از آن شامل اراضی کشاورزی است. خاک‌های منطقه موردمطالعه از مواد مادری حاصل از رسوبات آبرفتی یا آبرفتی- بادرفتی تشکیل شده‌اند که از مارن‌های گچی، سنگ‌آهک و ماسه‌سنگ منشأ گرفته‌اند. خاک‌های منطقه در دو رده انتی‌سولز[6] و اینسپتی-سولز[7] قرار دارند.

نمونه‌برداری و تجزیه‌های آزمایشگاهی

پس از بازدیدهای صحرایی، تعداد ۱۴۸ خاک‌رخ خاک بر اساس یک الگوی شبکه‌ای منظم با میانگین فاصله ۵۰۰ متر که در برخی مناطق بر اساس نظر کارشناس و شرایط محلی تغییر یافتند، حفر و مطالعه شدند. تشریح خاک‌رخ‌ها و ویژگی‌های مورفولوژیکی با روش شونبرگر و همکاران[8] انجام شد (۳۴). نمونه‌برداری از افق‌های ژنتیکی در تمام خاک‌رخ‌ها انجام شد. نمونه‌ها پس از هوا خشک‌کردن از الک ۲ میلی‌متری عبور داده شدند و ویژگی‌های قابلیت هدایت الکتریکی عصاره اشباع توسط دستگاه هدایت‌سنج در دمای ۲۵ درجه سلسیوس و پ‌هاش در عصاره‌های اشباع با استفاده از پ‌هاش متر (۳۱)، کربنات کلسیم معادل و گچ (۲۱)، بافت خاک با روش هیدرومتر (۱۴) و کربن آلی با استفاده از روش والکلی و بلک (۴۳) انجام شد. در ادامه بر اساس نتایج تشریح خاک‌رخ‌ها و تجزیه‌های فیزیکی و شیمیایی نمونه‌های خاک، خاک‌ها مطابق با سیستم جامع رده‌بندی خاک به روش آمریکایی (۳۸) طبقه‌بندی شدند.

1- Sharififar *et al.*
2- Mesic
3- Xeric
4- Hill lands
5- Piedmont plains
6- Entisols
7- Inceptisols
8- Schoeneberger *et al.*



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

روندنمای مدل‌سازی

فلوچارت انجام پژوهش شامل دو مرحله در شکل ۲ نشان داده شده است. مرحله اول با استفاده از نتایج به‌دست‌آمده از مطالعات صحرایی و آزمایشگاهی کلاس‌های خاک در نقاط مطالعاتی بر اساس تئوری اسکورپن[1] و با استفاده از الگوریتم‌های یادگیری ماشین جنگل تصادفی[2]، درخت تصمیم توسعه‌یافته[3] و رگرسیون لجستیک چندجمله‌ای[4] نقشه رقومی خاک تهیه و مورد ارزیابی قرار گرفت. در مرحله بعد به‌منظور کاهش اثر فراوانی کلاس‌های خاکی که ممکن است منجر به توزیع نامتعادل کلاس‌های خاک و پیش‌بینی نادرست نقشه خاک شود، سعی گردید از داده‌های متعادل برای تهیه نقشه کلاس خاک استفاده شود. برای رفع محدودیت یا چالش استفاده از داده‌های نامتعادل، ابتدا با استفاده از رویکرد نمونه‌گیری مجدد (کم نمونه‌گیری و بیش‌نمونه‌گیری) متعادل‌سازی داده‌ها انجام شد و در ادامه مدل‌سازی با الگوریتم‌های موردمطالعه مجدداً انجام و پیش‌بینی کلاس‌های خاک با استفاده از داده‌های نامتعادل و داده‌های متعادل مقایسه و ارزیابی شد.

متغیرهای کمکی

متغیرهای کمکی در واقع نماینده عوامل خاک‌سازی هستند (۲۶). عوامل خاک‌سازی با تأثیر بر روی فرآیندهای خاک‌سازی، تغییرپذیری خاک‌ها را رقم می‌زنند و استفاده از این عوامل و فرآیندهای خاک‌سازی می‌توانند به پیش‌بینی پراکنش خاک‌ها کمک نمایند. متغیرهای محیطی استفاده‌شده شامل اطلاعات نقشه ژئومورفولوژی، نقشه زمین‌شناسی، داده‌های سنجش‌ازدور و اطلاعات توپوگرافی هستند. برای این منظور نقشه زمین‌شناسی با مقیاس ۱:۲۵۰۰۰۰ از سازمان زمین‌شناسی کشور تهیه و در محیط سامانه اطلاعات جغرافیایی ArcGIS نسخه 10.7 زمین مرجع و رقومی‌سازی شد. هجده شاخص پستی‌وبلندی با استفاده از مدل رقومی ارتفاع با قدرت تفکیک مکانی ۳۰×۳۰ متر سنجنده استر، در محیط نرم‌افزار SAGA GIS (نسخه 7.9) استخراج شد. شاخص‌های

1- Scorpan
2- Random Forests (RF)
3- Boosted Regression Trees (BRT)
4- Multinomial Logistic Regression (MNLR)

سنجش‌ازدوری (۳۶ متغیر) با استفاده از تصاویر سنجنده (OLI/TIRS) ماهواره لندست ۸ با قدرت تفکیک مکانی ۳۰×۳۰ متر (USGS 2014) پس از اعمال تصحیحات رادیومتریکی و اتمسفری در محیط نرم‌افزار ENVI (نسخه 5.3) تهیه شد. نقشه ژئومورفولوژی بر اساس تلفیق لایه‌های اطلاعاتی واحدهای شکل زمین، مواد مادری به همراه تفسیر تصاویر ماهواره‌ای بر اساس رویکرد سلسله مراتبی ارائه‌شده توسط زینک (۴۴) تهیه گردید (شکل ۳). در ادامه بر اساس رویکرد تحلیل مؤلفه اصلی[1] در نرم‌افزار SPSS نسخه 26.0.0 و رتبه-بندی اهمیت نسبی مدل یادگیری ماشین از میان ۵۷ متغیر محیطی تولیدشده ۱۰ متغیر محیطی (جدول ۱) 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] پس از یکسان‌سازی مقیاس‌ها در محیط نرم‌افزار SAGA GIS به‌عنوان مؤثرترین متغیرهای محیطی برای پیش‌بینی کلاس‌های خاک و به‌عنوان ورودی مدل انتخاب گردید (۲۰).

مدل‌سازی مکانی

پس از آماده‌سازی داده‌ها و متغیرهای محیطی مؤثر در مدل‌سازی به‌عنوان متغیرهای وابسته و داده‌های مربوط به کلاس‌های خاک در منطقه موردمطالعه، اقدام به پیش‌بینی کلاس‌های خاک با استفاده از سه الگوریتم جنگل تصادفی (بسته Random Forest)، درخت تصمیم توسعه‌یافته (بسته C5.0) و رگرسیون لجستیک چندجمله-ای (بسته Caret) در محیط نرم‌افزار R_{Studio} (نسخه 2.3.492) اقدام شد.

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

جنگل تصادفی

مدل جنگل تصادفی یک تکنیک یادگیرنده فعال و توسعه‌یافته از مدل طبقه‌بندی و رگرسیون درختی 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 (۷). این روش یک روش یادگیری گروهی است و برای طبقه‌بندی با ساختن تعداد درختان زیاد عمل می‌نماید (۶). اساس روش‌های یادگیرنده گروهی این است که گروهی از یادگیرنده‌های ضعیف، مجموعه‌ای از یادگیرنده‌های قوی را تشکیل می‌دهند.

درخت تصمیم توسعه‌یافته

رگرسیون درختی توسعه‌یافته به‌عنوان یکی از الگوریتم‌های یادگیری ماشین ترکیبی از دو تکنیک آماری بوستینگ[9] و رگرسیون درختی است (۳). بوستینگ یک روش مرحله‌ای روبه‌جلو است که در آن مدل‌های درختی به‌صورت تکرارپذیر با زیرمجموعه‌ای از داده‌های آموزشی برازش داده می‌شوند. در برازش رگرسیون درختی توسعه‌یافته باید دو پارامتر نرخ یادگیری[10] و پیچیدگی درخت[11] مشخص گردند. نرخ یا مقدار یادگیری سهم هر درخت متوالی را در مدل نهایی تعیین می‌کند. پیچیدگی درخت اثرات اصلی یا اثرات متقابل بین متغیرها را نشان می‌دهد (۱۲).

رگرسیون لجستیک چندجمله‌ای

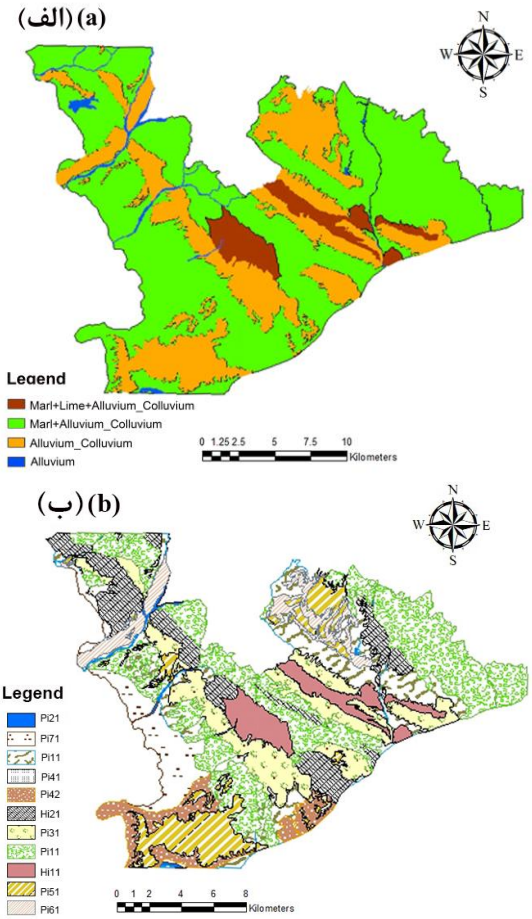
مدل رگرسیون لجستیک یک نوع مدل خطی تعمیم‌یافته است و برای مجموعه داده‌هایی مناسب است که متغیر وابسته به‌صورت دسته‌ای است. این مدل‌ها قادر به توصیف روابط بین مجموعه‌ای از متغیرهای پیش‌بینی کننده و یک متغیر وابسته دوبخشی است که دارای مقادیر

9- Boosting
10- Learning rate
11- Tree complexity

صفر یا چهار است (۱۷). به‌منظور گسترش رویکرد مدل رگرسیون لجستیک برای پیش‌بینی متغیرهای چندجمله‌ای، کمپن و همکاران[1] (۱۷) یک رویکرد رگرسیون لجستیک چندجمله‌ای را پیشنهاد کردند. در هر دو حالت مدل رگرسیون لجستیک چندجمله‌ای برای هر کلاس خاک در منطقه موردمطالعه توسعه یافت و روابط توپوگرافی و واحدهای طبقه‌بندی خاک از داده‌های خاک تعیین شدند.

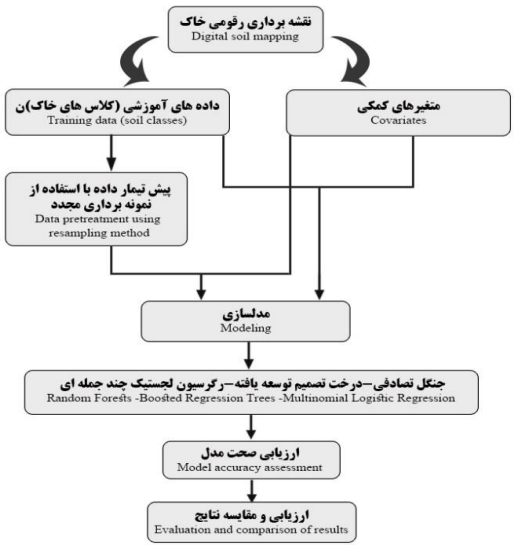
متعادل‌سازی داده با استفاده از روش نمونه-گیری مجدد از داده‌ها

برای بهبود داده‌های نامتعادل از دو عملکرد بهبود داده یعنی بیش نمونه‌گیری از کلاس‌های خاک اقلیت (تابع ubOver) و کم‌نمونه‌گیری از کلاس‌های خاک اکثریت (تابع ubUnder) از بسته نامتعادل (۱۰) در نرم‌افزار Rstudio استفاده شد. نمونه‌برداری کمتر برای کلاس‌های اکثریت A، B و D و نمونه‌برداری بیش‌ازحد برای کلاس‌های اقلیت E و F انجام گرفت (شکل ۴). کلاس‌های اکثریت کمتر از قبل و کلاس-های اقلیت تقریباً دو الی سه برابر نمونه‌برداری شدند تا توزیع به یک حد معمول نزدیک شود بدون اینکه نسبت اصلی کلاس‌های مختلف تغییر کند. روش کم-نمونه‌گیری به‌طور تصادفی تعداد مشاهده‌ها را در کلاس‌های اکثریت کاهش می‌دهد. درحالی‌که روش بیش‌نمونه‌گیری، مشاهده‌ها در کلاس‌های اقلیت خاک را تکرار می‌کند، به‌طوری‌که مشاهده‌ها اضافه‌شده از قبل در منطقه وجود داشته و هیچ داده نامربوطی اضافه نمی‌شود. در واقع هدف این است که توزیع داده‌های کلاس خاک متعادل شود (۲).



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

Figure (2)map of a) geology b) geomorphologi in study area



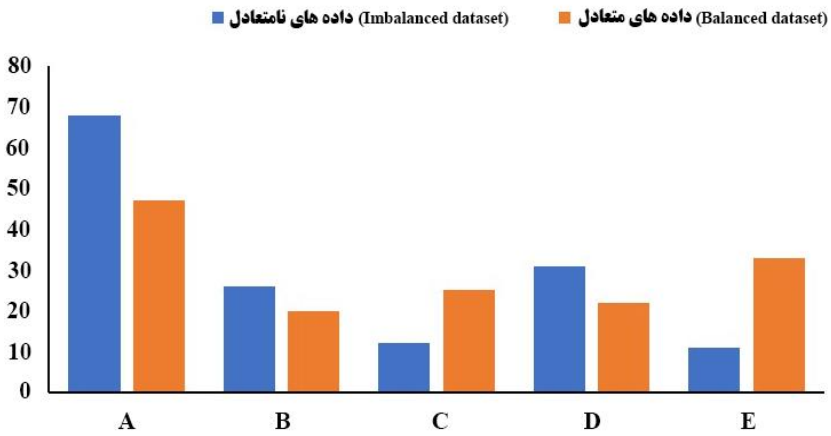
شکل (۳) فلوچارت مراحل انجام پژوهش

Figure (3) Research overview

1- Kempen *et al.*

جدول(۱) متغیرهای محیطی منتخب برای مدل‌سازی جهت پیش‌بینی کلاس‌های خاک
Table (1) Selected environmental covariates for modeling to predict soil classes

متغیر محیطی مورداستفاده Environmental Covariate used	ماهیت متغیر Covariate nature	مقیاس Scale	نام متغیر Covariate name	منبع Source
نقشه ژئومورفولوژی Geomorphology map	ژئومورفولوژی Geomorphology	۱:۵۰۰۰۰ 1:50000	نقشه ژئومورفولوژی Geomorphology map	Zinck et al (2016)
نقشه زمین‌شناسی Geology map	زمین‌شناسی Geology	۱:۲۵۰۰۰۰ 1:250000	نقشه زمین‌شناسی Geology map	Iran National Cartographic Center
مدل رقومی ارتفاع Digital elevation model	توپوگرافی Topography	۳۰ متر 30 meters	مدل رقومی ارتفاع DEM	ALOS PLASAR (2011)
			تجزیه‌وتحلیل سایه‌اندازی تپه‌ها Analytical Hill shading (AH)	Olaya (2004)
			طلوع خورشید Sunrise	Olaya (2004)
			عمق دره Valley Depth	Olaya (2004)
			شاخص طول در جهت شیب LS Factor	Olaya (2004)
			فاصله تا شبکه آبراهه Channel Network Distance (CHND)	Olaya (2004)
			شاخص خیسی توپوگرافی Topographic Wetness Index (TWI)	Olaya (2004)
			شاخص همواری بالای پشته با درجه تفکیک بالا Multi-resolution Ridge Top Flatness (MrRTF)	Olaya (2004)



شکل(٤) فراوانی کلاس‌های خاک قبل و بعد از نمونه‌برداری مجدد داده‌ها
Figure (4) Frequency of soil classes before and after data resampling

ارزیابی عملکرد مدل‌ها

اعتبار یک مدل به‌طور ساده بیان درصدی از پیش‌بینی‌های انجام‌شده توسط آن مدل است که با واقعیت موجود هماهنگی دارد. به‌منظور آموزش مدل‌ها مجموعه خاک‌رخ‌ها (متغیرهای محیطی و کلاس‌های خاک) به‌صورت تصادفی به دو مجموعه با نسبت چهار به یک تقسیم شدند. ۸۰ درصد داده‌ها برای آموزش مدل و ۲۰ درصد برای اعتبارسنجی استفاده شد. ارزیابی مدل‌ها با استفاده از شاخص‌های صحت کلی نقشه[1]، صحت تولیدکننده[2]، صحت کاربر[3] و ضریب کاپا[4] انجام شد (جنسن، ۱۹۹۶). صحت کلی نقشه از تقسیم تعداد کل پیکسل‌های که به‌درستی پیش‌بینی‌شده (مجموع قطر در ماتریس خطا) بر تعداد کل پیکسل‌های ماتریس خطا (N) به دست می‌آید (رابطه ۲).

(رابطه ۲): $OA = \sum_{i=1}^{k} X_{ii}/N$

در این رابطه Xii تعداد مشاهده‌ها در ردیف i و ستون i است، k تعداد سطرها (کلاس‌های خاک) در ماتریس خطااست.

شاخص کاپا معیاری برای مقایسه طبقه‌بندی مدل خودکار با طبقه‌بندی تصادفی است (رابطه ۳). این شاخص دارای مقداری بین صفر و یک است. اگر کاپا برابر با صفر باشد نشان‌دهنده طبقه‌بندی کاملاً تصادفی و مقدار منفی نشان‌دهنده خطا در طبقه‌بندی و اگر این مقدار برابر با یک باشد نشان‌دهنده طبقه‌بندی کاملاً صحیح است.

(رابطه ۳):

$$Kappa = N\sum_{i=1}^{k} X_{ii} - \sum_{i=1}^{k}(X_{i+} \times X_{+i})/N^2 - \sum_{i=1}^{k}(X_{i+} \times X_{+i})$$

در این رابطه Xi+ و X+j به ترتیب مجموع حاشیه‌ای برای ردیف i و ستون j هستند. مقادیر کاپا بیشتر از ۰/۸ نشان‌دهنده توافق یا دقت قوی بین نقشه طبقه‌بندی و اطلاعات مرجع زمینی است. مقادیر بین ۰/۴ و ۰/۸ نشان‌دهنده توافق متوسط و مقادیر کمتر از ۰/۴ نشان‌دهنده توافق ضعیف است (کنگلتون، ۱۹۹۱).

برای ارزیابی هر طبقه خاک، دو شاخص صحت تولیدکننده و صحت کاربر نیز محاسبه شد. دقت تولیدکننده از تقسیم تعداد کل پیکسل‌های صحیح یک کلاس بر تعداد کل پیکسل‌های آن کلاس از داده‌های مرجع زمین (کل ستون) محاسبه می‌شود و از رابطه ۴ به‌دست می‌آید.

(رابطه ۴): $PA = \frac{X_{jj}}{X_{+j}}$

در این رابطه Xjj و X+j به ترتیب مجموع حاشیه‌ای برای ردیف i و ستون j هستند.

صحت کاربر از تقسیم تعداد کل پیکسل‌های صحیح یک کلاس بر تعداد کل پیکسل‌هایی که واقعاً در آن دسته طبقه‌بندی شده‌اند (کل ردیف) محاسبه می‌شود و از رابطه ۵ به دست می‌آید.

(رابطه ۵) $UA = \frac{X_{ii}}{X_{+i}}$

در این رابطه Xii و X+i به ترتیب مجموع حاشیه‌ای برای ردیف i و ستون j هستند. دامنه تغییرات صحت تولیدکننده و صحت کاربر بین صفر و یک است و مقادیر بالاتر نشان‌دهنده عملکرد مناسب‌تر مدل است.

نتایج و بحث

خاک‌های منطقه در دو رده انتی‌سولز و اینسپتی‌سولز قرار دارند و در سطح تحت گروه شامل پنج کلاس تیپیک کلسی‌زرپتز[5]، تیپیک هاپلوزرپتز[6]، جیپسیک هاپلوزرپتز[7]، تیپیک زراورتنتز[8] و لیتیک زراورتنتز[9] شناسایی شدند. تحت گروه‌های خاک جیپسیک هاپلوزرپتز و لیتیک هاپلوزرپتز به ترتیب با فراوانی ۸/۱ و ۷/۳۴ درصد به‌عنوان کلاس‌های اقلیت و تحت گروه‌های تیپیک کلسی‌زرپتز، تیپیک هاپلوزرپتز، تیپیک زراورتنتز به ترتیب با فراوانی بیشتر از ۳۲/۴۳، ۱۷/۵۶ و ۲۰/۹۴ درصد از کل مشاهده‌های منطقه به‌عنوان کلاس‌های اکثریت در نظر گرفته شدند (جدول ۲).

1. Overall Accuracy, OA
2. Producer Accuracy, PA
3. Users Accuracy, UA
4. Kappa Index

5. Typic Calcixerepts
6. Typic Haploxerepts
7. Gypsic Haploxerepts
8. Typic Xerorthents
9. Lithic Xerorthents

نتایج مقادیر صحت‌سنجی پیش‌بینی مکانی هر یک از کلاس‌های خاک در شرایط معمول و بعد از رفع محدودیت داده‌های نامتعادل بر اساس چهار شاخص صحت کلی، شاخص کاپا، صحت کاربر و صحت تولیدکننده در جداول ۳ و ۴ نشان داده شده است. با توجه به نتایج ارائه‌شده در جدول (۳) قبل از متعادل‌سازی داده‌ها، مدل رگرسیون لجستیک چندجمله‌ای با شاخص کاپا برابر ۰/۴۱ و شاخص صحت کلی برابر ۶۶ درصد بالاترین دقت را نسبت به دو مدل جنگل تصادفی و درخت تصمیم توسعه‌یافته نشان داد، اما پس از متعادل‌سازی داده‌ها مدل درخت تصمیم توسعه‌یافته با شاخص کاپا برابر ۰/۶۴ و شاخص صحت کلی برابر ۷۵ درصد دقت بالاتری را در پیش‌بینی مکانی کلاس‌های خاک نشان داد (شکل ۵). نیستانی و همکاران[1] (۲۹) نشان دادند که مدل درخت تصمیم توسعه‌یافته با دقت کلی ۵۳ درصد و ضریب کاپا ۰/۳۹ بالاترین دقت برای برون‌یابی کلاس‌های خاک پس از متعادل‌سازی داده‌های خاک است. نتایج مطالعه ملاح و همکاران[2] (۲۴) نشان داد متعادل کردن داده‌های بافت خاک دقت کلی را از ۴۴ درصد به ۵۹ درصد و ضریب کاپا را از ۰/۳۰ به ۰/۵۲ بهبود بخشید.

در جدول (۴) نتایج دو شاخص صحت کاربر و صحت تولیدکننده برای کلاس‌های خاک در شرایط معمول و بعد از رفع محدودیت داده‌های نامتعادل در سطح زیرگروه ارائه شده است. مجموعه داده‌های اعتبارسنجی، صحت تولیدکننده و کاربر برای الگوریتم درخت تصمیم توسعه‌یافته تعداد بیشتری از کلاس‌ها با نتایج دقت نسبتاً بالاتر، در مقایسه با دو الگوریتم جنگل تصادفی و رگرسیون لجستیک چندجمله‌ای را نشان داد. زیرگروه‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنز که جزء کلاس‌های اقلیت محسوب می‌شوند هنگام استفاده از کلاس‌های نامتعادل توسط دو الگوریتم جنگل تصادفی و درخت تصمیم توسعه‌یافته پیش‌بینی‌نشده و به عبارتی حذف‌شده بودند و تنها الگوریتم رگرسیون لجستیک چندجمله‌ای قادر به پیش‌بینی مکانی این دو کلاس خاک بود اما پس از بهبود داده‌ها و بیش‌افزایی به تعداد این دو کلاس اقلیت پیش‌بینی این تحت گروه توسط دو الگوریتم جنگل تصادفی و درخت تصمیم توسعه‌یافته با صحت قابل قبولی افزایش نشان داد. این نتایج بیان می‌کند که تعداد نامتعادل مشاهده‌های کلاس‌های خاک می‌تواند تأثیر منفی بر نتایج اعتبارسنجی داشته باشد. شریفی‌فر و همکاران (۳۶) بهبود نتایج طبقه‌بندی با استفاده از مدل‌های میدان‌های تصادفی زنجیره مارکوف با روش نمونه‌برداری مجدد داده‌ها برای نقشه‌برداری کلاس‌های خاک را گزارش کردند. نیستانی و همکاران (۲۹) در مطالعه خود بیان کردند نمونه‌گیری بیش‌ازحد از کلاس خاک اقلیت منجر به افزایش دقت کلی در برخی مدل‌ها شده و کلاس‌های خاک اقلیت که در داده‌های نامتعادل نادیده گرفته شده است با نمونه‌برداری بیش‌ازحد، پیش‌بینی‌شده و در نقشه نهایی مشهود است. شریفی‌فر و همکاران (۳۵) در پژوهشی بیان کردند پس از بهبود داده‌ها، با استفاده بیش‌ازحد و کمتر از نمونه، همه مدل‌ها بهبود قابل‌توجهی در حفظ طبقات اقلیت، در ارزیابی را نشان دادند. تقی‌زاده مهرجردی و همکاران (۴۱) با مقایسه چند روش نمونه‌برداری مجدد از داده‌ها بیان کردند که تمامی روش‌های بهبود داده‌های نامتعادل سبب افزایش دقت پیش‌بینی مکانی کلاس‌های خاک پس از متعادل‌سازی داده‌ها می‌شوند. ملاح و همکاران (۲۴) اثرات تعادل و عدم تعادل کلاس‌های بافتی خاک از طریق پیش‌درمان داده‌ها را موردبررسی قرار داده و دریافتند درمان داده‌های بافت خاک با نمونه‌برداری کمتر و بیش‌ازحد، پیش‌بینی کلاس‌های بافت خاک‌های اقلیت را بهبود بخشید. محققین دیگر نیز در مطالعه خود گزارش کرده‌اند که تکنیک‌های نمونه‌گیری مجدد، یعنی نمونه‌گیری کمتر و بیش‌ازحد، دقت طبقه‌بندی را بهبود می‌بخشد (۲۲ و ۳۲).

شکل (۶) محتمل‌ترین نقشه‌های تولیدشده قبل و بعد از متعادل‌سازی داده‌ها با استفاده از الگوریتم‌های جنگل تصادفی، درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای را نشان می‌دهد. با توجه به نتایج صحت تولیدکننده و کاربر (جدول ۳) عملکرد هر دو کلاس اقلیت بهبود قابل‌توجهی را

1- Neyestani *et al.*
2- Mallah *et al.*

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 در الگوریتم‌های جنگل تصادفی و درخت تصمیم توسعه‌یافته حذف شده بودند، با دقت کاربر ۱۰۰ و ۷۸ در صد در الگوریتم جنگل تصادفی و ۶۰ و ۷۰ درصد در الگوریتم درخت تصمیم توسعه‌یافته با استفاده از داده‌های بهبودیافته، به‌خوبی پیش‌بینی شدند. کلاس‌های اقلیت دقت تولیدکننده ۷۵ و ۸۸ در صد در الگوریتم جنگل تصادفی و ۱۰۰ و ۷۸ درصد در الگوریتم درخت تصمیم توسعه‌یافته را در مقایسه با دقت صفر هنگام آموزش با استفاده از داده‌های نامتعادل نشان دادند. نتایج اعتبارسنجی الگوریتم رگرسیون لجستیک چندجمله‌ای نشان داد که علی‌رغم حفظ کلاس‌های اقلیت پس از متعادل‌سازی داده‌ها، پیش‌بینی کلاس‌های اقلیت با دقت کمتری انجام گرفته است. مطابق با نتایج اعتبارسنجی مدل درخت تصمیم توسعه‌یافته با بهبود دقت کاربر و تولیدکننده در چهار کلاس از پنج کلاس در سطح زیرگروه موفق به پیش‌بینی با دقت بالاتری پس از متعادل‌سازی داده‌ها شده و در واقع بالاترین احتمال پیش‌بینی برای همه کلاس‌ها در مقایسه با دیگر مدل‌ها را دارد. نتایج اعتبارسنجی مدل‌های مختلف در مطالعات مختلف نقشه‌برداری رقومی نشان می‌دهد که صرف‌نظر از نوع مدل به کار گرفته‌شده، نقشه‌های تهیه‌شده با استفاده از بهبود داده‌های نامتعادل دارای دقت بالاتری نسبت به نقشه‌های تهیه‌شده با داده‌های معمول بوده‌اند (۴۱ و ۳۶).

با توجه به اینکه پس از رفع محدودیت مدلی (درخت تصمیم توسعه‌یافته) که در شرایط معمول (با داده‌های نامتعادل) به‌صورت متوسط در نظر گرفته‌شده بود، در شرایط رفع محدودیت (با داده‌های متعادل) به‌عنوان بهترین مدل شناخته شد؛ می‌توان گفت توانایی مدل‌ها در شرایط داده‌های متعادل و نامتعادل یکسان نیست و استفاده از مدل‌های متنوع می‌تواند در هر دو روش شرایط معمول و رفع محدودیت منجر به انتخاب یک مدل ایده‌آل گردد. از طرفی توانایی بیشتر مدل‌های درختی توسعه‌یافته در پیش‌بینی مکانی کلاس‌های خاک در شرایط داده‌های متعادل را می‌توان به عملکرد و ساختار این مدل‌ها نسبت داد.

اهمیت نسبی متغیرهای محیطی در مدل درخت تصمیم توسعه یافته:

بر اساس نتایج برازش مدل‌های یادگیری ماشین بر روی کلاس‌های خاک، مدل درخت تصمیم بالاترین میزان صحت را برای پیش‌بینی زیرگروه‌های خاک نشان داد. در شکل (۷) نتایج اهمیت نسبی متغیرهای محیطی و در شکل (۸) پراکنش مکانی چند مورد از مهم‌ترین متغیرهای محیطی در مدل درخت تصمیم توسعه یافته نشان داده شده است. بر اساس نتایج از میان متغیرهای محیطی منتخب متغیرهای وابسته به توپوگرافی شامل عمق دره، شاخص خیسی توپوگرافی، مدل رقومی ارتفاع و فاصله تا شبکه آبراهه در بالاترین درجه اهمیت قرار دارند و سایر متغیرهای محیطی از اهمیت کمتری برخوردار هستند. مطابق با نتایج اینگونه استنباط می‌شود که توپوگرافی مهم‌ترین عامل خاک‌سازی در منطقه موردمطالعه است. در همین راستا موسوی و همکاران (۲۸) پارامترهای توپوگرافی را به‌عنوان مهم‌ترین پیشران‌های محیطی برای مدل-سازی کلاس‌های خاک گزارش نمودند. عباس‌زاده افشار و همکاران[1] (۱) بیان کردند متغیرهای محیطی مستخرج از مدل رقومی ارتفاع به دلیل داشتن همبستگی بالا با کلاس‌های خاک موجب ارتقاء صحت نقشه‌های رقومی خاک تهیه شده است.

1- Abbaszadeh Afshar *et al.*

جدول(۲) تعداد مشاهدات کلاس‌های خاک در سطح زیرگروه

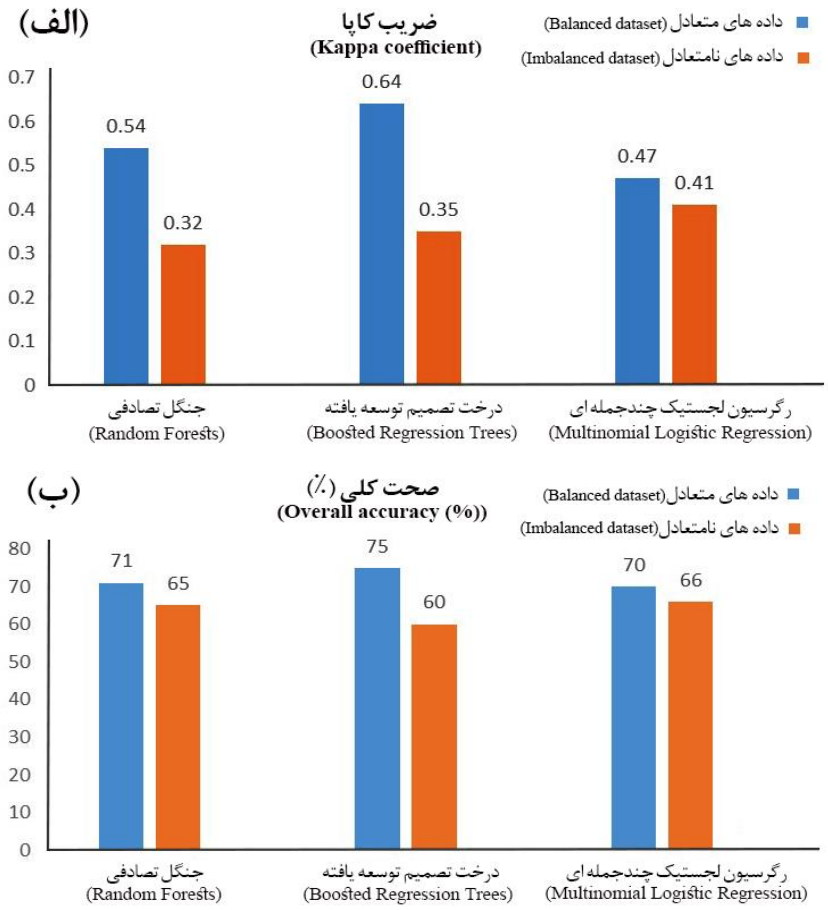
Table (2) The number of observations of soil classes at the subgroup level

کد کلاس Class code	تحت گروه‌های خاک Soil subgroups	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

جدول(۳) صحت پیش‌بینی سطح تاکسونومیک زیرگروه قبل و بعد از بهبود یا درمان داده‌ها توسط الگوریتم‌های یادگیرنده

Table(3) Prediction accuracy of the taxonomic level of the subgroup before and after improving or data treatment by learning algorithms

مدل‌های یادگیری ماشین Machine learning models	شاخص‌های صحت‌سنجی Validation indicators			
	ضریب کاپا Kappa coefficient		صحت کلی (٪) Overall accuracy (%)	
	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset
جنگل تصادفی RF	0.32	0.54	65	71
درخت تصمیم توسعه‌یافته BRT	0.35	0.64	60	75
رگرسیون لجستیک چندجمله‌ای MNLR	0.41	0.47	66	70



شکل (۵) نمودار الف) ضریب کاپا و ب) صحت کلی توسط الگوریتم‌های یادگیری ماشین قبل و بعد متعادل‌سازی داده‌ها

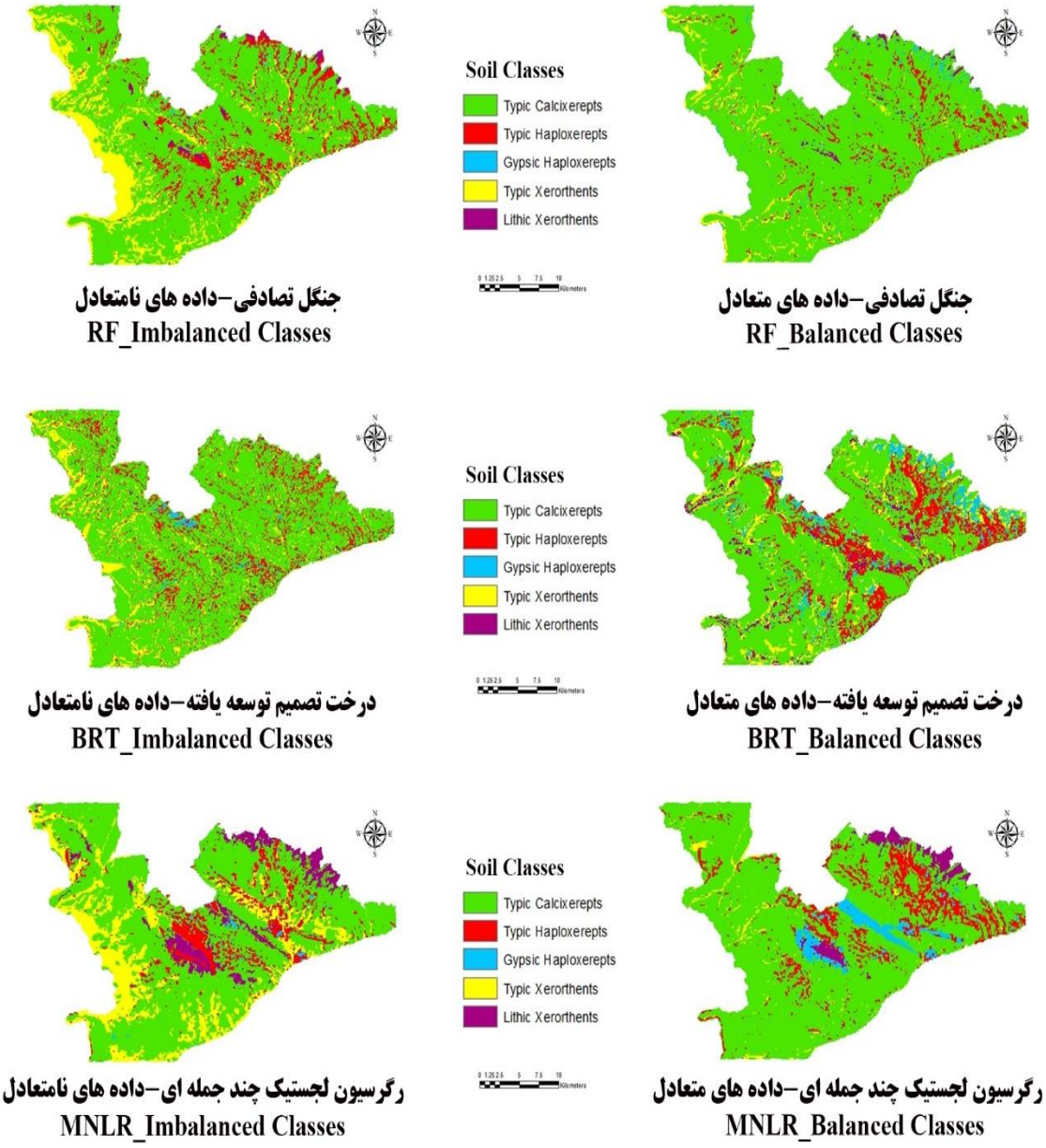
Figure (5) Chart a) Kappa coefficient and b) overall accuracy by machine learning algorithms before and after data balancing

جدول (٤) صحت تولیدکننده و کاربر برای کلاس‌های خاک در سطح زیرگروه قبل و بعد از بهبود یا درمان داده بر اساس مدل‌های برازش داده‌شده

Table (4) Producer and User accuracy for soil classes at the subgroup level before and after data treatment based on the fitted models

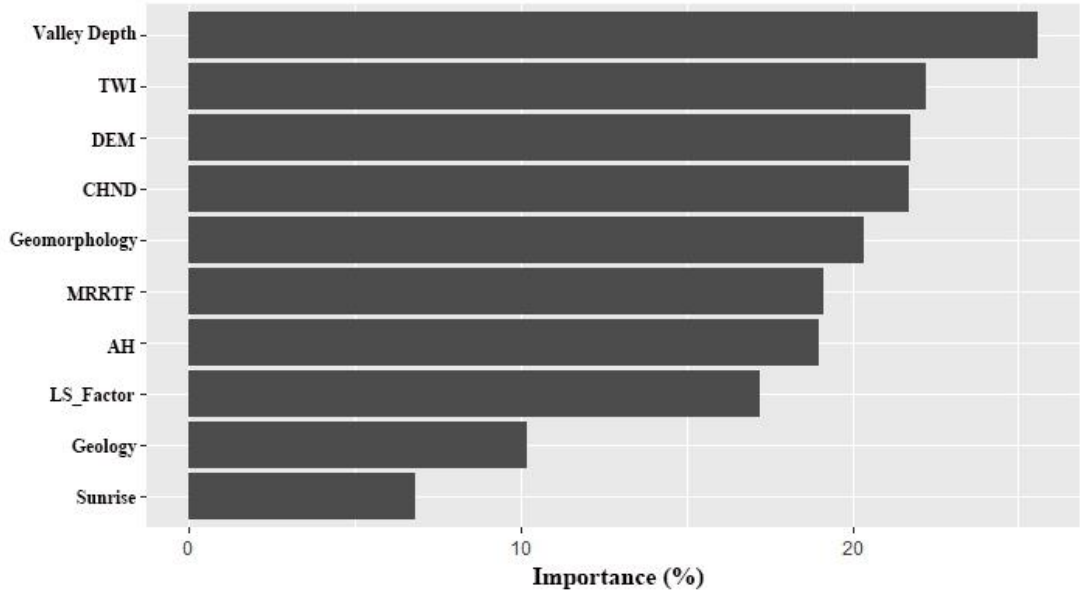
قابلیت اطمینان Validation	صحت کاربر (٪) User accuracy (%)						صحت تولیدکننده (٪) Producer accuracy (%)					
مدل‌های یادگیری ماشین Machine learning models	جنگل تصادفی RF		درخت تصمیم توسعه‌یافته BRT		رگرسیون لجستیک چندجمله‌ای MNLR		جنگل تصادفی RF		درخت تصمیم توسعه‌یافته BRT		رگرسیون لجستیک چندجمله‌ای MNLR	
کلاس خاک سطح زیرگروه Subgroup of soil	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset	داده‌های متعادل Balanced dataset
Typic Calcixerepts	61	75	62	94	64	72	85	85	80	74	94	90
Typic Haploxerepts	100	34	67	50	100	75	50	25	50	67	67	100
Gypsic Haploxerepts	NaN	100	NaN	60	20	17	0	75	0	100	34	30
Typic Xerorthents	65	34	40	50	67	100	34	34	40	50	40	25
Lithic Xerorthents	NaN	78	NaN	70	100	50	0	88	0	78	100	75

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



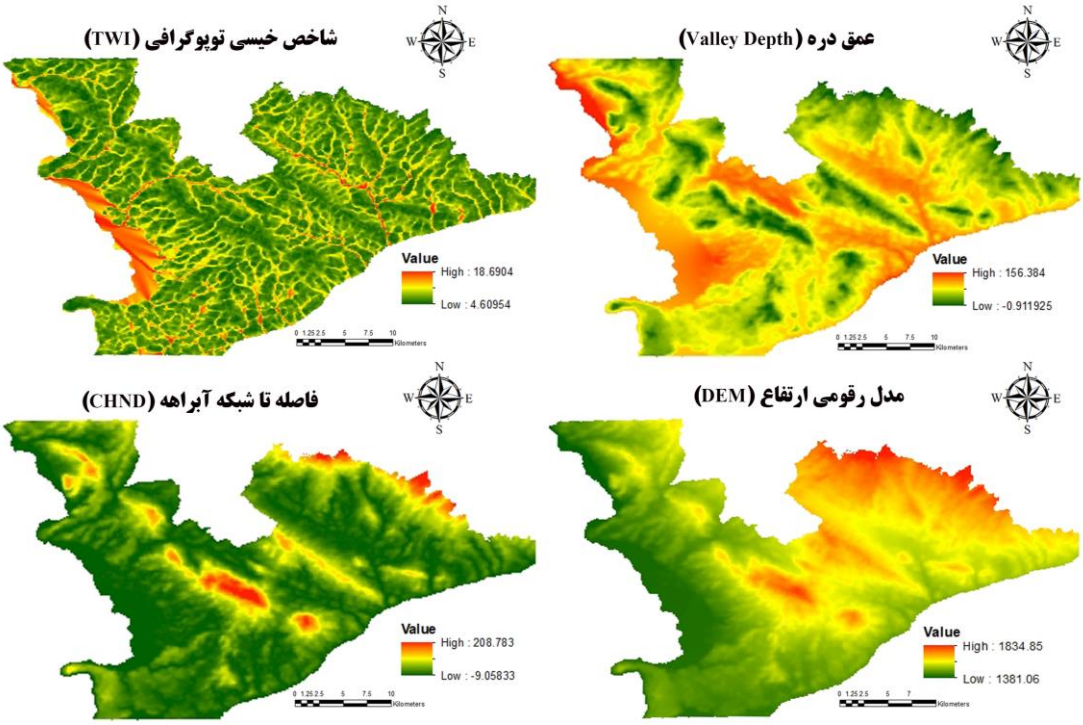
شکل(٦) نقشه‌های تولیدشده توسط الگوریتم‌های یادگیری ماشین قبل و بعد متعادل‌سازی داده‌ها

Figure (6) Maps produced by machine learning algorithms before and after data balancing



شکل(۷) اهمیت نسبی متغیرهای محیطی پیش‌بینی کننده کلاس‌های خاک در سطح زیرگروه بر اساس مدل درخت تصمیم توسعه یافته

Figure (7) The relative importance of environmental variables predicting soil classes at the subgroup level in the BRT model



شکل (۸) نقشه مهم‌ترین متغیرهای محیطی پیش‌بینی کننده کلاس خاک در مدل درخت تصمیم توسعه یافته

Figure (8) the map of the most important environmental variables predicting the soil class in the BRT model

نتیجه‌گیری

استفاده از الگوریتم‌های معمول برای نقشه‌برداری رقومی کلاس‌های خاک بدون توجه به تعداد کلاس‌های مشاهده‌شده نامتعادل خاک در یک منطقه می‌تواند منجر به کم‌توجهی مدل به کلاس‌های کم تکرار یا از دست دادن کلاس‌های اقلیت و تخمین بیش‌ازحد مدل از کلاس‌های اکثریت یا پر تکرار شود. این پژوهش با بررسی کارآمدی روش‌های پیش درمانی یا بهبود به‌منظور متعادل‌سازی داده‌ها با استفاده از روش نمونه‌گیری مجدد نشان داد که کاربرد روش‌های بهبود داده‌های نامتعادل برای تهیه نقشه رقومی خاک منجر به تهیه نقشه‌های دقیق‌تر و با جزئیات بیشتر شده است. در رابطه با کارایی مدل‌ها بر اساس نتایج این مطالعه پس از پردازش داده‌ها، با روش بیش‌نمونه‌گیری و کم‌نمونه‌گیری، الگوریتم‌های جنگل تصادفی و درخت تصمیم توسعه‌یافته بهبود قابل‌توجهی در حفظ کلاس‌های کم تعداد، نشان دادند در شرایطی که مدل‌سازی با روش‌های متداول باعث حذف یا نادیده گرفته شدن کلاس‌های کم تعداد و پیش‌بینی نادرست نقشه خاک شده بود. همچنین بر اساس نتایج اعتبارسنجی الگوریتم درخت تصمیم توسعه‌یافته بالاترین دقت کلی و ضریب کاپا را در مقایسه با دو الگوریتم جنگل تصادفی و رگرسیون لجستیک چندجمله‌ای نشان داد؛ بنابراین می‌توان گفت روش نمونه‌گیری مجدد داده‌ها می‌تواند یک‌راه حل مفید برای مقابله با داده‌های نامتعادل کلاس خاک به‌ویژه در مدل درخت تصمیم توسعه‌یافته برای تولید نقشه‌های رقومی خاک باشد. با توجه به اینکه مطالعات در این زمینه بسیار محدود بوده و منابع کمی در اختیار است این پژوهش می‌تواند بینشی جدید در مورد نقشه‌برداری رقومی خاک با تعداد مشاهده‌های نامتعادل و متعادل در مقابل پژوهشگران قرار دهد.

References

1. Abbaszadeh Afshar, F., and Ayubi, Sh., and Jafari, A., 2017. Spatial prediction of large soil groups using regression models and decision tree in the southeast region of Iran. Agricultural Engineering (Agricultural Scientific Journal), 41(2): 133-146.

2. 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), pp.238-251.

3. Abeare, S., 2009. Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico lonline [sic] fishery. Louisiana State University and Agricultural & Mechanical College.

4. Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B. and Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PloS one, 9(8), p.e105519.

5. Alibeigi, M., Hashemi, S. and Hamzeh, A., 2012. DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets. Data & Knowledge Engineering, 81, pp.67-103.

6. Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

7. Breiman, L. and Cutler, A., 2004. Random Forests. Department of Statistics, University of Berkeley.

8. Caubet, M., Dobarco, M.R., Arrouays, D., Minasny, B. and Saby, N.P., 2019. Merging country, continental and global predictions of soil texture: Lessons from ensemble modelling in France. Geoderma, 337, pp.99-110.

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

10. Dominati, E., Patterson, M. and Mackay, A., 2010. A framework for classifying and quantifying the natural capital and ecosystem services of soils. Ecological economics, 69(9), pp.1858-1868.

11. Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. Journal of animal ecology, 77(4), pp.802-813.

12. Fernández, A., del Jesus, M.J. and 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), pp.9805-9812.

13. Gee, G.W. and Or, D., 2002. 2.4 Particle-size analysis. Methods of soil analysis: Part 4 physical methods, 5, pp.255-293.

14. 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, pp.62-77.

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

16. Kempen, B., Brus, D.J., Heuvelink, G.B. and Stoorvogel, J.J., 2009. Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma, 151(3-4), pp.311-326.

17. Kleinbaum, A.M., 2018. Reorganization and tie decay choices. Management Science, 64(5), pp.2219-2237.

18. Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), pp.221-232.

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

20. Loeppert, R.H. and Suarez, D.L., 1996. Carbonate and gypsum. Methods of soil analysis: Part 3 chemical methods, 5, pp.437-474.

21. Loyola-González, O., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A. and García-Borroto, M., 2016. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. Neurocomputing, 175, pp.935-947.

22. Ma, Y., Minasny, B., Malone, B.P. and Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). European Journal of Soil Science, 70(2), pp.216-235.

23. Mallah, S., Delsouz Khaki, B., Davatgar, N., Scholten, T., Amirian-Chakan, A., Emadi, M., Kerry, R., Mosavi, A.H. and Taghizadeh-Mehrjardi, R., 2022. Predicting Soil Textural Classes Using Random Forest Models: Learning from Imbalanced Dataset. Agronomy, 12(11), p.2613.

24. Malone, B.P., Minasny, B., McBratney, A.B., Malone, B.P., Minasny, B. and McBratney, A.B., 2017. Digital Soil Assessments. Using R for Digital Soil Mapping, pp.245-260.

25. McBratney, A.B., Santos, M.M. and Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1-2), pp.3-52.

26. 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, p.103437.

27. Mousavi, S.R., Sarmadian, F., and 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), pp.2525-2538.

28. 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, p.e00422.

29. Pozzolo, A.D., Caelen, O. and Bontempi, G., 2015. Unbalanced: Racing for unbalanced methods selection. R package version, 2.

30. Ramentol, E., Vluymans, S., Verbiest, N., Caballero, Y., Bello, R., Cornelis, C. and Herrera, F., 2014. IFROWANN: imbalanced fuzzy-rough ordered weighted average nearest neighbor classification. IEEE Transactions on Fuzzy Systems, 23(5), pp.1622-1637.

31. Richards, A. L. (ed). 1954. Diagnosis and improvement of saline and alkaline soils. US Salinity Laboratory Staff. USDA. Handbook, No. 60, Washington DC. USA.

32. Sáez, J.A., Krawczyk, B. and Woźniak, M., 2016. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognition, 57, pp.164-178.

33. Sasada, T., Liu, Z., Baba, T., Hatano, K. and Kimura, Y., 2020. A resampling method for imbalanced datasets considering noise and overlap. Procedia Computer Science, 176, pp.420-429.

34. Schoeneberger, P. J., Wysocki, D. A., and Benham, E. C. (Eds.). 2012. Field book for describing and sampling soils. Government Printing Office.

35. Sharififar, A., Sarmadian, F., Malone, B.P. and Minasny, B., 2019. Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma, 350, pp.84-92.

36. Sharififar, A., Sarmadian, F. and Minasny, B., 2019. Mapping imbalanced soil classes using Markov chain random fields models treated with data resampling technique. Computers and Electronics in Agriculture, 159, pp.110-118.

37. Soil and Water Research Institute. 2010. Site Selection, Soil Survey and Land Evaluation for Development of Orchards in Zanjan Province, Iran.

38. Soil Survey Staff. 2014. Keys to soil taxonomy, 12th edition. USDA Natural Resources Conservation Service.

39. Statistical Yearbook of Zanjan Province. 2019. Land and Climate, National Statistics Organization.

40. Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B. and Zhou, Y., 2015. A novel ensemble method for classifying imbalanced data. Pattern Recognition, 48(5), pp.1623-1637.

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

42. 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, p.114552.

43. Walkley, A. and 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), pp.29-38.

44. 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.