Journal of Range & Watershed Management
https://jrwm.ut.ac.ir/

Volume 76 (2)
Online ISSN: 2423-7795

Investigating the effectiveness of resampling algorithms in improving the classification of unbalanced data in digital soil mapping

Fatemeh Ebrahimi Meymand[1*] | Hasan Ramezanpour[1] | Nafiseh Yaghmaeian[1] | Kamran Eftekhari[2]

1. Soil Science Department, College of Agriculture, University of Guilan, Rasht, Iran.
2. Soil and Water Research Institute, Agriculture Research, Education and Extension Organization (AREEO), Karaj, Iran.
E-mail: meymand1949@gmail.com

Article Info

Article type:
Research Article

Article history:
Received: 28 Jan. 2023
Received in Revised from: 13 Mar. 2023
Accepted: 18 Mar. 2023
Published online: 22 Aug. 20233

Keywords:
Machine learning,
Oversampling,
R software,
soil map,
under sampling.

Abstract

In recent years, the use of digital soil mapping (DSM) based on machine learning algorithms for preparing soil maps has become widespread. These algorithms predict soil classes by modeling the relationships between them and environmental variables. However, one challenge with this method is the imbalanced distribution of soil in the landscape. This imbalance leads to overfitting and underfitting of classes, resulting in reduced accuracy of the models used. This study evaluates the ability of two machine learning algorithms, random forests and support vector machines, for digitally mapping soil classes using an imbalanced dataset. The study focuses on 95 soil profile classes at the family level, covering 4000 hectares of land in the Honam sub-basin, Lorestan province. To address the issue of imbalance in soil classes, six datasets were examined. These datasets include the original soil dataset and five datasets created through various resampling approaches. These approaches include two manual classifications and three methods of over-sampling, under-sampling, and Synthetic Minority Over-Sampling Techniques in the R software. The results demonstrate that despite the low overall accuracy, the geographical distribution of soils with high frequency in the study area in the digital soil map obtained from the random forest and the original dataset, as well as the Synthetic Minority Over-Sampling Technique, corresponds significantly with the conventional soil map of the study area. Therefore, the low number of observations for other soil classes, resulting in incorrect training of models, can be considered one of the main reasons for the low accuracy of the models used.

Cite this article: Ebrahimi Meymand, F., Ramezanpour, H., Yaghmaeian, N., Eftekhari, K. (2023). Investigating the effectiveness of resampling algorithms in improving the classification of unbalanced data in digital soil mapping. *Journal of Range & Watershed Management*, 76 (1), 159-176. DOI: http//doi.org/10.22059/jrwm.2023.354333.1692

Publisher: University of Tehran Press

نشریه مرتع و آبخیزداری
دوره ۷۶ (۲)
شاپا الکترونیکی: ۷۷۹۵-۲۴۲۳

بررسی کارایی رویکردهای باز نمونه‌گیری در بهبود کلاس‌بندی داده‌های نامتوازن در نقشه‌برداری رقومی خاک

فاطمه ابراهیمی میمند[۱*] | حسن رمضانپور[۱] | نفیسه یغمائیان[۱] | کامران افتخاری[۲]

۱. گروه علوم خاک، دانشکده کشاورزی، دانشگاه گیلان، رشت، ایران
۲. مؤسسه تحقیقات خاک و آب، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران
رایانامه: meymand1949@gmail.com

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخ دریافت: ۱۴۰۱/۱۱/۰۸
تاریخ بازنگری: ۱۴۰۱/۱۲/۲۲
تاریخ پذیرش: ۱۴۰۱/۱۲/۲۷
تاریخ انتشار: ۱۴۰۲/۰۵/۳۱

کلیدواژه‌ها:
بیش‌نمونه‌گیری،
کم‌نمونه‌گیری،
نرم‌افزار R،
نقشه خاک،
یادگیری ماشین.

چکیده

در سال‌های اخیر، استفاده از روش‌های نقشه‌برداری رقومی مبتنی بر الگوریتم‌های یادگیری ماشین باهدف تهیه نقشه کلاس‌های خاک بطور گسترده‌ای توسعه یافته است. اساس این روش‌ها پیش‌بینی کلاس‌ها یا ویژگی‌های خاک به کمک مدل‌سازی روابط بین آن‌ها و متغیرهای محیطی به عنوان نمایندگان عوامل خاک‌سازی، می‌باشد. ماهیت نامتوازن توزیع خاک‌ها در طبیعت که منجر به بیش‌برازش کلاس‌های با فراوانی زیاد و کم برازش کلاس‌های با فراوانی کم و درنتیجه کاهش دقت فرآیند نقشه‌برداری خاک شده، از چالش‌های موجود در این روش می‌باشد. بنابراین، پژوهش حاضر باهدف ارزیابی توانایی دو الگوریتم جنگل تصادفی و ماشین‌بردارپشتیبان در نقشه‌برداری‌رقومی کلاس‌های فامیل خاک با توزیع نامتوازن، حاصل از ۹۵ خاک‌رخ مطالعاتی در ۴۰۰۰ هکتار از اراضی زیرحوضه هنام، استان لرستان انجام گرفت. در این مطالعه موضوع عدم توازن در فراوانی کلاس‌های خاک با استفاده از ۶ مجموعه داده، ازجمله مجموعه داده‌های اصلی و پنج مجموعه داده ایجادشده توسط چندین رویکرد نمونه‌گیری مجدد از داده‌های اصلی، شامل دو رویکرد طبقه‌بندی دستی و سه الگوریتم بیش‌نمونه‌گیری و کم‌نمونه‌گیری و بیش‌نمونه‌گیری اقلیت مصنوعی در محیط نرم افزار R موردبررسی قرار گرفت. نتایج نشان داد علیرغم مقایر پائین آماره‌های اعتبارسنجی، شباهت گسترش خاک‌های با فراوانی زیاد در منطقه مطالعاتی در نقشه‌های حاصل از مدل جنگل تصادفی و مجموعه داده‌های اصلی و همچنین الگوریتم بیش‌نمونه‌گیری اقلیت مصنوعی با نقشه خاک تهیه‌شده به روش مرسوم قابل‌توجه می‌باشد. بنابراین فراوانی کم سایر کلاس‌های خاک و در نتیجه آن عدم آموزش درست مدل‌ها برای آن‌ها را می‌توان یکی از دلایل اصلی صحت‌کلی کم مدل‌های به‌کاررفته دانست.

استناد: ابراهیمی میمند؛ فاطمه، رمضانپور؛ حسن، یغمائیان؛ نفیسه، افتخاری؛ کامران (۱۴۰۲). بررسی کارایی رویکردهای باز نمونه‌گیری در بهبود کلاس‌بندی داده‌های نامتوازن در نقشه‌برداری رقومی خاک. *نشریه مرتع و آبخیزداری*، ۷۶(۲)، ۱۷۶-۱۵۹.

DOI: http//doi.org/10.22059/jrwm.2023.354333.1692

ناشر : انتشارات دانشگاه تهران.

۱. مقدمه

دستیابی به اطلاعات دقیق از وضعیت خاک‌ها و نحوه پراکنش مکانی آن‌ها از نیازهای اساسی کاربرانی است که در زمینه‌های مختلف کشاورزی، منابع طبیعی و محیطزیست مشغول فعالیت می‌باشند. این‌گونه اطلاعات به‌صورت مرسوم در قالب نقشه‌های پلی گونی خاک‌شناسی با صرف وقت و هزینه زیاد تهیه و ارائه می‌شوند. امروزه متناسب با پیشرفت‌های مستمر در محاسبات، فناوری‌های سنجش‌ازدور و نزدیک و سیستم‌های اطلاعات جغرافیایی، استفاده از روش‌های نوین نقشه‌برداری که با نام کلی نقشه‌برداری رقومی خاک[1] شناخته می‌شوند به‌عنوان یکی از زیرشاخه‌های موفق در علم پدومتری مطرح شده است(Minasny, and McBratney, 2016; pahlavan-Rad et al., 2016).

مفهوم نقشه‌برداری رقومی خاک اولین بار توسط مک برانتی و همکاران[2] (۲۰۰۳) ارائه گردید. در واقع نقشه‌برداری رقومی خاک شامل استفاده ازروش‌ها و مدل‌های مختلف، جهت ایجاد ارتباط بین توزیع خاک (کلاس‌ها یا ویژگی‌های خاک) و داده‌هایی که به‌آسانی و باقیمت ارزان از طریق روش‌های سنجش‌ازدور، تصاویر ماهواره‌ای و داده‌های ژئومورفومتری به دست می‌آیند و تحت عنوان متغیرهای کمکی محیطی نامیده می‌شوند، هست. تکنیک‌های نقشه‌برداری رقومی خاک بر روی محاسبات رقومی استوار هستند، اما اساس آن‌ها بر روی معادلات تشکیل خاک قرار دارد (pahlavan-Rad et al., 2016).

در سال‌های اخیر، استفاده از رویکرد نقشه‌برداری رقومی و تکنیک‌های مختلف یادگیری ماشین جهت تهیه نقشه‌های کلاس خاک در سطوح مختلف رده‌بندی رواج گسترده‌ای یافته است (Brungard et al., 2016; Hengl et al., 2017; Ma et al., 2020; Massawe et al., 2018; Ramcharan et al., 2018). اگرچه، دقت مدل‌های به‌کاررفته در تولید نقشه‌های خاک در این مطالعات به تعداد کلاس‌های خاک و توزیع فراوانی آن‌ها که تابعی از پیچیدگی محیطی و ماهیت سیستم طبقه‌بندی خاک در منطقه موردمطالعه است، بستگی دارد (Sharififar et al., 2019). بنابراین افزایش سطح رده‌بندی در مطالعات خاکشناسی که منجر به افزایش تعداد کلاس‌های خاک و به‌تبع آن کاهش تعداد مشاهدات در هر گروه می‌شود، باعث دشوار شدن آموزش مدل و کاهش صحت پیش‌بینی‌ها می‌شود. زیرا مدل نمی‌تواند ارتباط بین پراکنش خاک‌هایی با فراوانی کم و پارامترهای محیطی را به‌خوبی تشخیص دهد (Mosleh et al., 2016; Mousavi et al., 2020; Nazari et al., 2020).

این موضوع که اکثر الگوریتم‌های طبقه‌بندی یادگیری ماشین به عدم تعادل در کلاس‌های پیش‌بینی حساس هستند و در یک مجموعه داده نامتعادل، مدل پیش‌بینی‌ها را به سمت کلاس‌های با فراوانی بیشتر هدایت می‌کند و باعث حذف کلاس‌های با فراوانی کم می‌شود، در مطالعات مختلف مورد تائید قرارگرفته است (Haixian et al., 2017; Jafari et al., 2013; Sharififar et al., 2019; Thomas, 1996). این‌گونه داده‌ها که تحت عنوان داده‌های نامتوازن شناخته می‌شوند، یکی از مشکلات موجود در طبقه‌بندی داده‌ها می‌باشد (Haixiang et al., 2017; Neyestani et al., 2021). بنابراین، به نظر می‌رسد که با توجه به هدف موردنظر و شرایط منطقه مطالعاتی، برای آنکه بتوان در سطوح پایین رده‌بندی به صحت قابل قبولی در تخمین دست یافت می‌توان راهکارهای مختلفی را مدنظر قرار داد (Zhu et al., 2017). یکی از این راهکارها افزایش تعداد مشاهدات در کلاس‌هایی است که فراوانی آن‌ها کم می‌باشد. البته این موضوع، مستلزم صرف وقت و هزینه زیاد می‌باشد. راهکار دیگر کاهش تعداد کلاس‌های خاک با در نظر گرفتن خاک‌های مشابه از لحاظ مدیریتی است، این روش به‌شرط آنکه با توجه به منطقه مورد مطالعه بتولند تعداد کلاس‌های خاک را تا اندازه زیادی کاهش دهد؛ می‌تولند در افزایش صحت پیش‌بینی مؤثر باشد (Mosleh et al., 2016). آدهیکاری و همکاران[3] (۲۰۱۴) گزارش نمودند که با در نظر گرفتن خاک‌های مشابه، صحت مدل برای پیش‌بینی کلاس‌های خاک، ۱۶ درصد افزایش یافته است. از ترکیب کلاس‌های خاکی که فراوانی کمی دارند با یکدیگر و قرار دادن آن‌ها تحت عنوان یک گروه، مثلاً سایر خاک‌ها می‌توان به‌عنوان روش دیگر نام برد (Brungard et al., 2015).

از دیگر تکنیک‌های قابل‌قبول و پرکاربرد برای مواجهه با داده‌های نامتوازن بر روی الگوریتم‌های یادگیری ماشین، روش‌های

[1] Digital Soil Mapping (DSM)
[2] McBratney et al
[3] Adhikari et al

بازنمونه‌گیری[1] است (Abdi and Hashemi, 2016; Chawla et al., 2002). این روش‌ها در مطالعات مختلف تهیه نقشه رقومی خاک‌ها برای افزایش صحت پیش‌بینی الگوریتم‌های یادگیری ماشین استفاده شده‌اند (Neyestani et al., 2021; Sharififar et al., 2019; Taghizadeh-Mehrjardi et al., 2015).

بنابراین، پژوهش حاضر باهدف مقایسه دو الگوریتم یادگیری ماشین جنگل تصادفی و ماشین‌بردار پشتیبان، در مدل‌سازی و تهیه نقشه رقومی کلاس‌های خاک در سطح فامیل سامانه رده‌بندی آمریکایی، انجام گردید. همچنین کارایی چند رویکردها بازنمونه‌گیری در بهبود کلاس‌بندی داده‌های نامتوازن و در نتیجه کاهش اثرات این‌گونه داده‌ها بر نتایج مدل‌سازی مورد بررسی قرار گرفت.

۲. مواد و روش‌ها

۲-۱. منطقه مطالعاتی

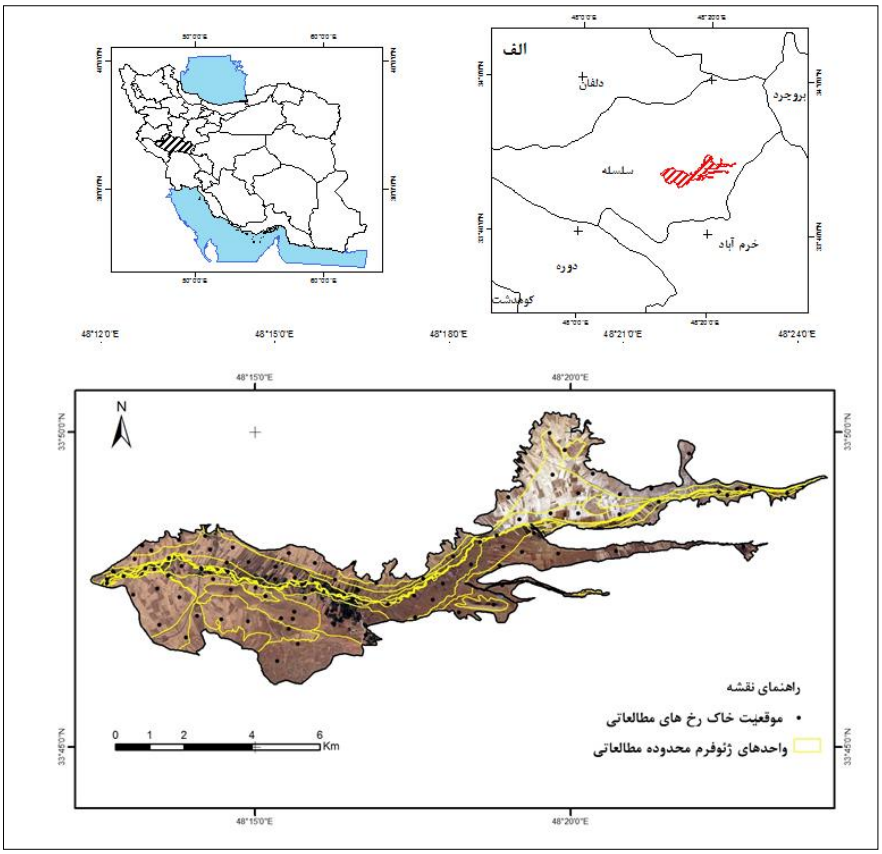
منطقه مورد مطالعه در بخشی از اراضی زیرحوضه هنام در استان لرستان، جنوب شهر الشتر، بین طول جغرافیایی "۰۰'۱۲°۴۸ تا "۰۰ '۲۴° ۴۸ شرقی و عرض‌های جغرافیایی"۰۰' ۴۷° ۳۳ تا "۰۰' ۴۹° ۳۳ شمالی، در مساحتی حدود ۴۰۰۰ هکتار واقع شده و دارای رژیم رطوبتی زریک و رژیم حرارتی مزیک است (Banai, 1998). کاربری اصلی اراضی در این زیر حوضه کشت آبی (گندم، جو، کلزا، انواع لوبیا، علوفه و چغندرقند) و دیم (گندم، جو و نخود) گیاهان زراعی می‌باشد. مساحت کمی از اراضی نیز به باغات میوه (گردو و میوه‌های سردسیری) اختصاص داده شده است. مهم‌ترین منبع آب سطحی زیرحوضه، رودخانه هنام است. میانگین بارندگی سالانه منطقه ۵۵۴ میلی‌متر و میانگین سالیانه دمای هوا ۸/۸ درجه سانتی‌گراد می‌باشد (https://www.accuweather.com/fa/ir/alashtar.). رسوبات کواترنر که از سنگ‌های رسوبی و سایر تشکیلات زمین‌شناسی ارتفاعات مجاور منشاء گرفته‌اند بخش وسیعی از منطقه مطالعاتی را در برگرفته‌اند (https://gsi.ir/fa/map.). این منطقه در دو زمین‌نمای دشت دامنه‌ای و دره واقع شده است و زمین‌نمای دشت دامنه‌ای بخش اعظم منطقه را در بر گرفته است. متوسط ارتفاع منطقه ۱۷۰۰ متر نسبت به سطح دریای آزاد و دامنه تغییرات شیب صفر تا ۲۴ درصد است (Eftekhari, 2019).

۲-۲. تعیین نقاط مطالعاتی

جهت تعیین نقاط نمونه‌برداری سعی گردید تا با استفاده از نقشه اولیه ژئوفرم‌ها، با مقیاس ۱:۲۰۰۰۰ و در سطح زمین‌ریخت تهیه شده بود، به‌گونه‌ای عمل گردد که با انعطاف‌پذیری در انتخاب نقاط، تا حد امکان در اشکال مختلف اراضی که نمونه تیپیک خاک‌های محدوده می‌باشند، حداقل یک نقطه مشاهداتی جهت حفر خاکرخ‌های مطالعاتی انتخاب گردد. بعضی مناطق به لحاظ پیچیدگی پدیده‌های موجود، به صورت تفصیلی‌تر (فواصل کمتر نقاط مطالعاتی) مورد مطالعه قرار گرفت و بالعکس برای مناطق همگن‌تر فواصل بیشتری برای شبکه‌های نمونه‌برداری طراحی گردید. در شکل شماره ۱ موقعیت خاکرخ‌های مطالعاتی ارائه شده است.

طی مطالعات میدانی در منطقه مورد مطالعه ۹۵خاک‌رخ حفر و تشریح گردید و از تمامی افق‌های ژنتیکی قابل شناسایی نمونه‌برداری انجام شد و پس از هوا خشک کردن و عبور از الک ۲ میلیمتری نمونه‌ها برای آزمایشات فیزیکوشیمیایی معمول به آزمایشگاه منتقل و بافت خاک و اجزای آن شامل رس، سیلت و شن خاک به روش هیدرومتری (Jackson, 1973)، اسیدیته در گل اشباع با دستگاه pH متر، هدایت الکتریکی عصاره اشباع توسط دستگاه هدایت‌سنج در دمای ۲۵ درجه سانتی گراد، کربنات کلسیم معادل به روش تیتراسیون برگشتی با اسید کلریدریک یک نرمال، مواد آلی به روش سوزاندن تر با بیکرومات پتاسیم در مجاورت اسیدسولفوریک غلیظ اندازه‌گیری شد (Thomas, 1996). درپایان، تمامی خاکرخ‌های مطالعاتی بر اساس مشاهدات میدانی و نتایج تجزیه آزمایشگاهی برطبق سامانه رده‌بندی آمریکایی تا سطح فامیل طبقه‌بندی شدند (Soil Survey Staff, 2014).

[1] Resampling



شکل ۱. موقعیت جغرافیایی منطقه مورد مطالعه در شهرستان سلسله، استان لرستان و موقعیت خاکرخ‌های مطالعاتی

۲-۳. نقشه‌برداری رقومی خاک

مفهوم نقشه‌برداری رقومی خاک اولین بار توسط مک برانتی و همکاران (۲۰۰۳) ارائه گردید. تکنیک‌های نقشه‌برداری رقومی خاک بر روی محاسبات رقومی استوار هستند، اما اساس آن‌ها بر روی معادلات تشکیل خاک قرار دارد. در واقع نقشه‌برداری رقومی خاک شامل استفاده از روش‌ها و مدل‌های مختلف، جهت ایجاد ارتباط بین توزیع خاک (کلاس‌ها یا ویژگی‌های خاک) و داده‌هایی که به‌آسانی از طریق روش‌های سنجش‌ازدور، تصاویر و عکس‌های ماهواره‌ای و داده‌های ژئومورفومتری به دست می‌آیند و تحت عنوان متغیرهای کمکی محیطی نامیده می‌شوند، می‌باشد. بنابراین یکی از مهم‌ترین گام‌ها در پروژه‌های نقشه‌برداری رقومی، تهیه متغیرهای کمکی است. متغیرهای کمکی محیطی درواقع نماینده عوامل خاک‌سازی هستند. اکثر متغیرهای کمکی با استفاده از مدل‌های رقومی ارتفاع[1]، داده‌های طیفی سنجش‌ازدور و نقشه‌های حاصل از مطالعات گذشته به‌دست می‌آید. در این مطالعه برخی از مشتقات پستی و بلندی از قبیل ارتفاع[2]، درجه شیب[3]، جهت شیب[4]، انحنای شیب افقی[5]، انحنای شیب نیمرخی[6]، شاخص خیسی[7] و شاخص حمل رسوب[8]، فاصله تا

[1] Digital Elevation Model (DEM)
[2] Elevation
[3] Slope
[4] Aspect
[5] Plan Curvature
[6] Profile Curvature
[7] Topographical Wetness Index
[8] Sediment Transport Capacity

نزدیک‌ترین نهر یا رودخانه[۱]، عمق دره[۲]، ارتفاع نرمال شده[۳]، ارتفاع شیب[۴]، فنوتیپ‌های ژئومورفولوژی[۵] (Jasiewicz and Stepinski, 2013)، بر اساس روش‌های مبتنی بر تجزیه پستی و بلندی، از لایه مدل رقومی ارتفاع منطقه مطالعاتی مستخرج از ماهواره SRTM که باقدرت تفکیک ۳۰ متر به‌صورت رایگان در دسترس است، با استفاده از نرم‌افزار SAGA-GIS استخراج شدند.

در این مطالعه با استفاده محیط نرم‌افزار تحت وب گوگل ارث انجین[۶] تصاویر سری زمانی ماهواره سنتینل ۲، مربوط به ماه‌های خرداد تا شهریور ماه سال ۱۳۹۸ (مطابق با سال انجام مطالعات صحرایی) با قدرت تفکیک مکانی ۱۰ متر استخراج و پس از انجام پیش‌پردازش‌های لازم، میانگین باندهای طیفی این تصاویر جهت محاسبه شاخص‌های پوشش گیاهی و شاخص‌های طیفی مربوط مواد مادری و خاک استفاده شد. تعدادی از متغیرهای کمکی مورداستفاده در این مطالعه در شکل ۲ ارائه شده است. در پایان تمامی پارامترهای محیطی به نقشه‌های شبکه‌ای (رستری) با اندازه پیکسل ۳۰ متر تبدیل و پایگاه داده‌هایی مشتمل بر داده‌های خاکی (کلاس خاک در سطح فامیل) و داده‌های متغیرهای محیطی تشکیل شد.

۲-۴. مجموعه داده‌های مورداستفاده در آموزش مدل

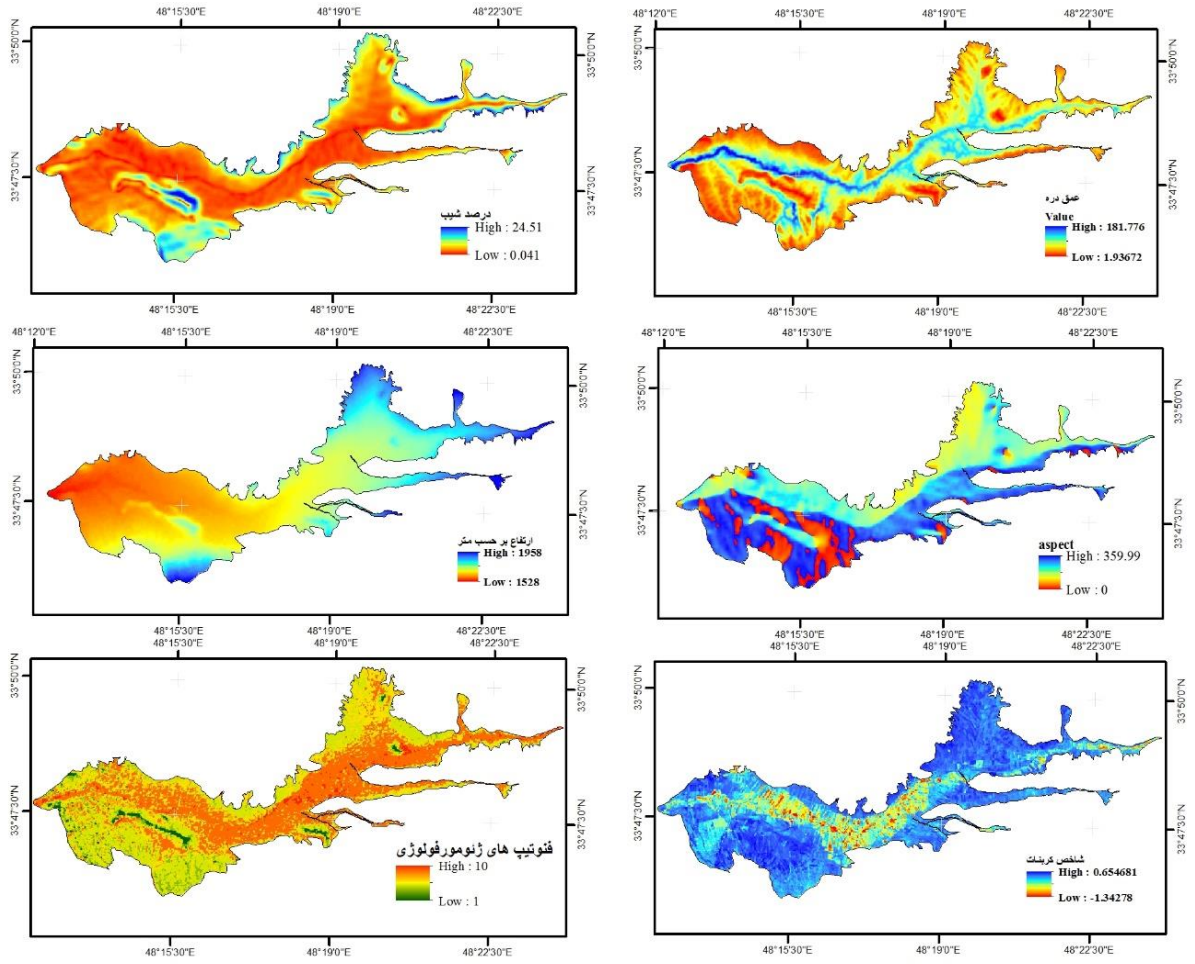
با توجه به نامتوازن بودن کلاس‌های خاک در خاکرخ‌های مطالعاتی و حساس بودن اکثر الگوریتم‌های طبقه‌بندی یادگیری ماشین به عدم تعادل در کلاس‌های پیش‌بینی از چندین رویکرد برای کاهش اثرات داده‌های نامتوازن بر نتایج مدل‌سازی استفاده شد و در پایان صحت عمومی پیش‌بینی‌ها توسط رویکردهای مورداستفاده با روش‌های استاندارد مدل‌سازی بدون انجام هیچ‌گونه تغییری بر روی کلاس‌های خاک مقایسه گردید. در اولین روش پیشنهادی کلاس‌های خاکی با فراوانی کم با یکدیگر ترکیب و تحت عنوان گروه سایر خاک‌ها نام‌گذاری شد (رویکرد اول). درروش پیشنهادی دیگرکلاس‌های خاک با در نظر گرفتن خاک‌های مشابه ازلحاظ مدیریتی مورد طبقه‌بندی قرار گرفتند و درنتیجه آن خاکرخ‌های مطالعاتی در۴ گروه طبقه‌بندی شدند (رویکرد دوم).

در این رویکرد ۱۷ کلاس خاک موجود در منطقه مطالعاتی با در نظر گرفتن خاک‌های مشابه از لحاظ مدیریتی مورد طبقه‌بندی قرارگرفته و درنتیجه آن خاکرخ‌های مطالعاتی در ۴ گروه جدید طبقه‌بندی شدند. اینسپتی‌سول‌های دارای ویژگی‌های ورتیک با خاک تحت‌الارضی خیلی سنگین در کلاس مدیریتی MO1، اینسپتی‌سول‌های فاقد ویژگی‌های ورتیک و دارای خاک تحت‌الارضی خیلی سنگین در کلاس مدیریتیMO2، خاک‌های حاوی بیش از ۳۵ درصد حجمی ذرات بزرگتراز ۲ میلیمتر در کلاس مدیریتی MO3 و خاک‌های دارای بافت تحت‌الارضی متوسط، در کلاس مدیریتی MO4 قرار گرفتند. طبقه‌بندی کلاس‌های خاک در این دو رویکرد به‌صورت دستی انجام گرفت.

در رویکرد سوم از سه الگوریتم بازنمونه‌گیری برای بهبود کلاس‌بندی داده‌های نامتوازن در آموزش مدل استفاده شد. ساده‌ترین راه برای اصلاح مجموعه داده‌های نامتعادل، متعادل کردن آن‌ها با نمونه‌برداری بیش‌ازحد از نمونه‌های کلاس اقلیت[۷] یا کم‌نمونه‌گیری[۸] از نمونه‌های کلاس اکثریت است. روش دیگر تکنیک نمونه‌برداری بیش‌ازحد اقلیت مصنوعی[۹] است. این روش توسط چالا و همکاران[۱۰] (۲۰۰۲) به‌عنوان یک روش نمونه‌برداری پیشنهاد شد و به‌عنوان یک پیش‌پردازش بر روی‌داده‌ها استفاده شد. با استفاده از این الگوریتم می‌توان تعداد نمونه‌های کلاس اقلیت را توسط نمونه‌های مصنوعی تولیدشده افزایش داد.

---

[1] Distance to Nearest Stream
[2] Valley depth (vd)
[3] Normalized Hight (NH)
[4] Slope Hight (SLH)
[5] Geomorphologic phonotypes (Geomorphons)
[6] Google earth engine
[7] Oversampling
[8] Undersampling
[9] Synthetic Minority Over-Sampling Technique (SMOTE)
[10] Chawla et al



شکل ۲. برخی از داده‌های محیطی مهم شناخته شده در مدل‌های پیش‌بینی

۲-۵. مدل‌سازی

در این مرحله از تحقیق مدل‌های جنگل تصادفی [1] و ماشین بردار پشتیبان [2] برای پیش‌بینی کلاس‌های خاک در سطح فامیل (Soil Survey Staff, 2014) استفاده شدند. جنگل تصادفی یک الگوریتم یادگیری ماشین متشکل از چندین درخت تصمیم است که به دلیل سادگی و قابلیت استفاده از پرکاربردترین الگوریتم‌های مورداستفاده در بسیاری از مطالعات نقشه‌برداری رقومی خاک محسوب می‌شود (Mosleh et al., 2016; Mirakzehi et al., 2018). مدل ماشین بردار پشتیبان یکی دیگر از الگوریتم‌های یادگیری ماشین مورداستفاده در نقشه‌برداری رقومی خاک است که داده‌ها را به‌صورت نقاطی در فضا در نظر گرفته و آن‌ها را با استفاده از یک خط یا صفحه از هم جدا می‌کند. این جداسازی به‌گونه‌ای است که نقاط داده‌ای که در یک‌طرف خط یا صفحه هستند مشابه به هم و در یک گروه قرار می‌گیرند (Kovačević et al., 2010). این دو مدل در محیط نرم‌افزار R و با استفاده از بسته نرم‌افزاری caret اجرا شدند. سهم هرکدام از متغیرهای محیطی در پیش‌بینی کلاس‌های خاک و تشخیص مهم‌ترین پارامترهای اثرگذار در مدل جنگل تصادفی با استفاده از

[1] Random Forest (RF)
[2] Support vector machine (SVM)

تابع var imp در نرم افزار R صـــورت گرفت. این تابع از معیاری به نام شـــاخص جینی[1] برای انتخاب مهم‌ترین تغییرها کمک می‌گیرد (Martinez and Redondo, 2020).

سهم پارامترهای محیطی در پیش‌بینی کلاس‌های خاک در مدل SVM نیز با استفاده از شاخص کارائی کلی[2] (TP) ارائه‌شده توسط باقری و همکاران (۲۰۱۶) به دست آمد.

$$TP=\frac{TA}{OA_{train}}$$

$$TA=\frac{OA_{train}.n_{train}+OA_{test}.n_{test}}{n_{tot}}$$

در این معادله TA: نمایانگر درصـــد برآوردهای درســـت در آموزش و آزمون مدل، OA_{train} : صـــحت کلی مدل آموزش، OA_{test}: صـــحت کلی مدل اعتبارســـنجی، n_{train}: تعداد داده‌های آموزش و n_{test}: تعداد داده‌های اعتبارســـنجی مدل و n_{total}: تعداد کل داده‌ها می‌باشد. هرچه مقدار نسبت اخیر به عدد یک نزدیک باشد، نشان‌دهنده برآورد بهتر مدل است.

پس از انتخاب پارامترهای مهم برای مدل‌ســـازی در مراحل جداگانه کلاس‌های خاک در ســـطح فامیل به همراه پارامترهای محیطی برای مدل‌ها تعریف و بر اســـاس ارتباط کلاس‌های خاک با پارامترهای محیطی، پیش‌بینی‌ها صـــورت گرفت. این مدل‌های بر روی کلاس‌های خاک حاصـــل از ۶۰ خاک رخ مطالعاتی آموزش داده شـــدند. اطلاعات مربوط به کلاس‌های خاک ۳۵ خاک رخ مطالعاتی نیز برای اعتبارسنجی مدل‌ها کنار گذاشته شد.

۲-۶. ارزیابی صحت مدل‌ها

اعتبار یک مدل به‌طور ساده بیان درصدی از پیش‌بینی‌های انجام شده توسط آن مدل است که با واقعیت موجود هماهنگی دارد. ارزیابی صحت پیش‌بینی کلاس‌های خاک با استفاده از ماتریس خطا (Congalton,1991) و شاخص‌های صحت عمومی (OA)، شاخص کاپا (K) با استفاده از روابط زیر صورت پذیرفت.

$$OA=\sum_{i=1}^{n}\frac{X_{ii}}{N}$$

$$K=N\sum_{i=1}^{n}xX_{ii}-\sum_{i=1}^{n}(X_{io}\times X_{oi})/N^2-\sum_{i=1}^{n}(X_{io}-X_{oi})$$

در این روابط، n تعداد سطر یا ستون‌های ماتریس، Xii کلاس‌هایی که به‌درستی پیش‌بینی شده‌اند، Xio تعداد کل سطرها، Xoi تعداد کل ستون‌ها و N تعداد کل مشاهدات می‌باشد. بالاترین میزان صحت عمومی و شاخص کاپا، به‌عنوان معیار انتخاب بهترین مدل برای پیش‌بینی کلاس‌های خاک، در نظر گرفته شد.

علاوه بر ارزیابی صـحت پیش‌بینی کلاس‌ها با شـاخص‌های مذکور، نقشـه‌های رقومی تهیه شـده با نقشـه کلاس‌های خاک منطقه مطالعاتی که به روش مرسوم مطالعات خاکشناسی تهیه شده بود، مورد مقایسه قرار گرفتند.

[1] Gini Index
[2] Total Proficiency (TP)

۳. یافته‌های پژوهش

نتایج حاصل از تشریح خاکرخ‌های مطالعاتی حاکی از جوان بودن خاک‌ها و قرار داشتن آن‌ها در دو رده انتی سول و اینسپتی‌سول است. افق سطحی اکریک و افق تحت‌الارضی کلسیک مهم‌ترین افق‌های پدوژنیکی تشکیل شده در منطقه می‌باشند، که در شرایط ترمودینامیکی و تحت تاثیر فعالیت فاکتورهای خاکسازی حاکم بر منطقه مطالعاتی تشکیل شده‌اند. جدول ۱ کلاس‌های خاک مربوط به ۹۵ خاک رخ مطالعاتی و تعداد هر کلاس را نشان می‌دهد. همان‌طور که در جدول ارائه‌شده است از مجموع ۱۷ کلاس خاک شناسایی‌شده در منطقه مطالعاتی، حدود۸۰ درصد خاک‌رخ‌های مطالعاتی در پنج فامیل خاک (A-E) طبقه‌بندی شده‌اند و بقیه کلاس‌های خاک (F-Q) دارای فراوانی بسیار کم می‌باشند. این موضوع به‌خوبی بیانگر توزیع نامتوازن کلاس‌های خاک شناسایی شده در منطقه مطالعاتی است.

جدول ۱. طبقه‌بندی خاکرخ‌های مطالعاتی در سطح فامیل خاک (۳۴) و فراوانی هر کلاس

ردیف	رده‌بندی خاک درسطح فامیل خاک	کد کلاس‌های خاک*	تعدادخاکرخ مطالعاتی	درصد خاکرخ‌ها
۱	Fine, mixed, mesic, Vertic Haploxerepts	A	۲۳	۲۴/۲۱
۲	Fine, mixed, mesic, Typic Haploxerepts	B	۲۲	۲۳/۱۶
۳	Fine, mixed, mesic, Typic Calcixerepts	C	۲۱	۲۲/۱۱
۴	Fine, mixed, mesic, Vertic Calcixerepts	D	۹	۹/۴۷
۵	Fine, carbonatic, mesic, Typic Calcixerepts	E	۵	۵/۲۶
۶	Fine loamy, carbonatic, mesic, shallow Petrocalcic Calcixerepts	F	۲	۲/۱۱
۷	Clay skeletal, mixed, mesic, Typic Calcixerepts	G	۲	۲/۱۱
۸	Fragmental, mixed, mesic, Typic Xerofluvents	H	۲	۲/۱۱
۹	Coarse loamy, mixed, mesic, Typic Xerofluvents	I	۱	۱/۰۵
۱۰	Fine loamy, carbonatic, mesic, Typic Calcixerepts	J	۱	۱/۰۵
۱۱	Fine loamy, mixed, mesic, Typic Calcixerepts	K	۱	۱/۰۵
۱۲	Fine loamy, mixed, mesic, Typic Haploxerepts	L	۱	۱/۰۵
۱۳	Loamy skeletal, mixed, calcareous, mesic, Typic Xerofluvents	M	۱	۱/۰۵
۱۴	Clayey skeletal, mixed, mesic, Fluventic Haploxerepts	N	۱	۱/۰۵
۱۵	Loamy skeletal, mixed, mesic, Fluventic Haploxerepts	O	۱	۱/۰۵
۱۶	Loamy skeletal, mixed, mesic, Typic Calcixerepts	P	۱	۱/۰۵
۱۷	Sandy skeletal, mixed, mesic, Typic Xerofluvents	Q	۱	۱/۰۵
	جمع کل		۹۵	۱۰۰

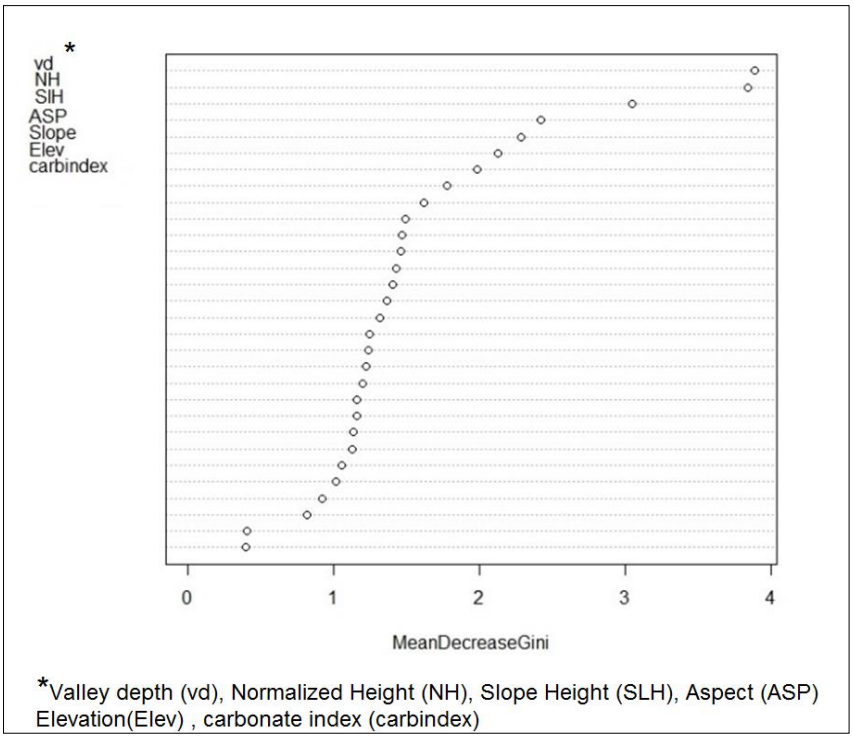
* کدهای مورد استفاده در مقاله برای سهولت در نشان دادن کلاس‌های خاک

مجموعه داده‌های مورد استفاده در آموزش مدل ازجمله مجموعه داده‌های اصلی خاک و مجموعه داده‌های حاصل از ترکیب کلاس‌های خاکی با فراوانی کم با یکدیگر و نام‌گذاری آن‌ها تحت عنوان سایر خاک‌ها در جدول ۲ ارائه‌شده است.

قابلیت اعتماد به نتایج پیش‌بینی به میزان زیادی متأثر از توانایی پارامترهای محیطی در بیان تغییرات فاکتور موردبررسی می‌باشد (Mosleh et al., 2016). نتایج حاکی از آن است که از بین ۳۰ متغیر محیطی موردبررسی متغیرهای عمق دره، ارتفاع نرمال شده ارتفاع شیب، جهت شیب، درجه شیب، ارتفاع از سطح دریا و شاخص کربنات به ترتیب مهم‌ترین متغیرها در پیش‌بینی کلاس خاک توسط مدل جنگل تصادفی (شکل۳) و متغیرهای جهت شیب، ارتفاع شیب، ارتفاع از سطح دریا، فنوتیپ های ژئومورفولوژی، عمق دره، شاخص کربنات و درجه شیب به ترتیب مهم‌ترین متغیرها در پیش‌بینی کلاس خاک توسط مدل SVM شناخته شدند (جدول ۳).

جدول ۲. طبقه‌بندی خاکرخ‌های مطالعاتی مورد استفاده در آموزش مدل

ردیف	مجموعه داده‌های آموزش مدل با داده‌های اصلی		مجموعه داده‌های آموزش مدل با ترکیب کلاس‌های دارای فراوانی کم	
	کلاس طبقه‌بندی	فراوانی هر کلاس	کلاس طبقه‌بندی	فراوانی هر کلاس
۱	A	۱۷	A	۱۷
۲	B	۱۴	B	۱۴
۳	C	۱۴	C	۱۴
۴	D	۶	D	۶
۵	E	۳	E	۳
۶	F	۱	other soils	۶
۷	G	۱		
۸	H	۱		
۹	M	۱		
۱۰	N	۱		
۱۱	Q	۱		
	جمع کلاس‌ها			۶۰

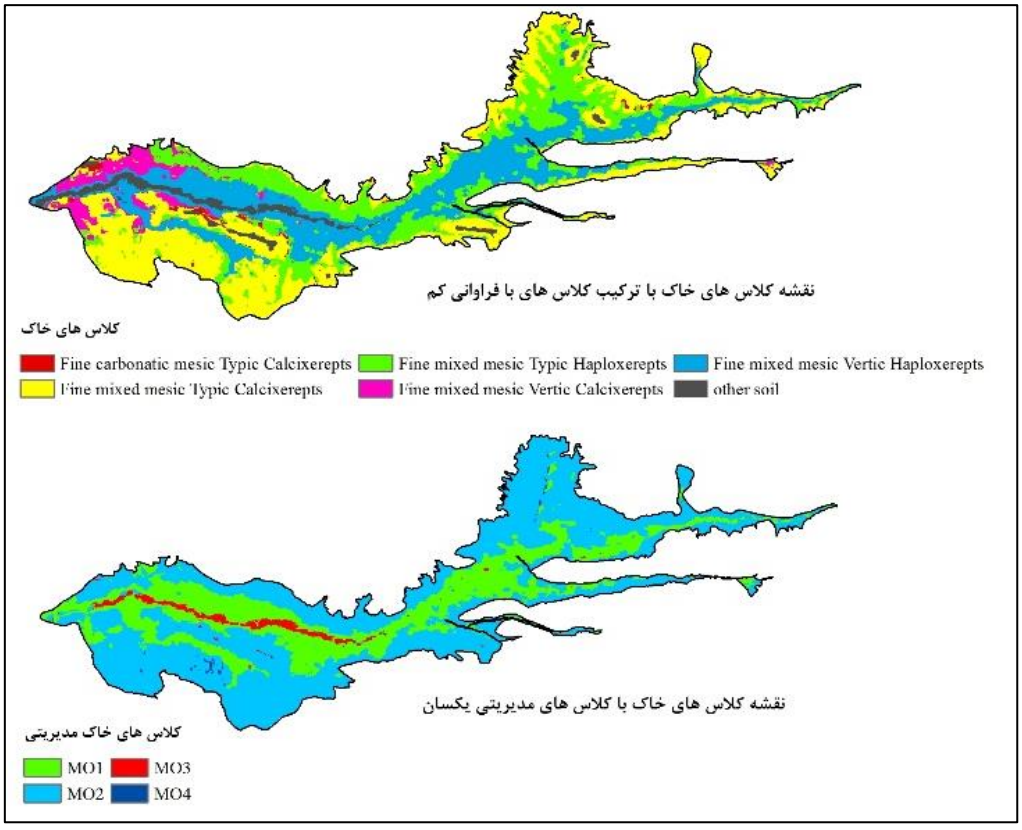


شکل۳. اهمیت پارامترهای محیطی در پیش‌بینی کلاس‌های خاک در مدل جنگل تصادفی

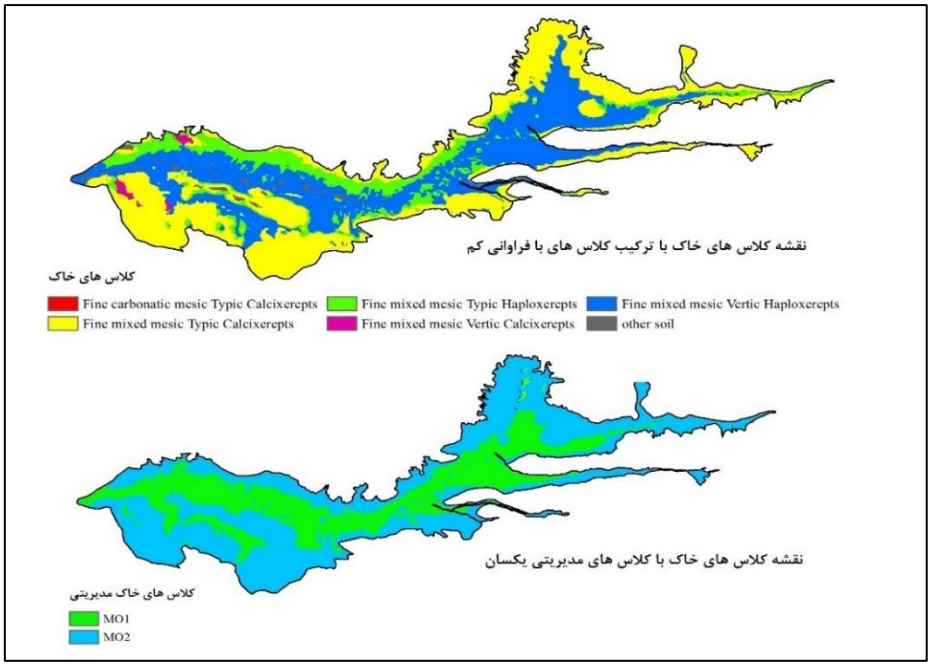
جدول ۳. اهمیت پارامترهای محیطی در پیش‌بینی کلاس‌های خاک در مدل SVM

پارامتر محیطی	OA_{train}	N_{train}	OA_{test}	N_{test}	TA	TP
جهت شیب	۰/۴۴		۰/۳۷		۰/۴۱	۰/۹۴
ارتفاع شیب	۰/۳۶		۰/۲۵		۰/۳۲	۰/۸۹
ارتفاع از سطح دریا	۰/۲۵		۰/۱۷		۰/۲۲	۰/۸۸
فنوتیپ‌های ژئومورفولوژی	۰/۴۳	۶۰	۰/۲۸	۳۵	۰/۳۷	۰/۸۷
عمق دره	۰/۴۸		۰/۳۱		۰/۴۲	۰/۸۷
شاخص کربنات	۰/۴۵		۰/۲۸		۰/۳۹	۰/۸۶
درجه شیب	۰/۳۷		۰/۲۲		۰/۳۱	۰/۸۵

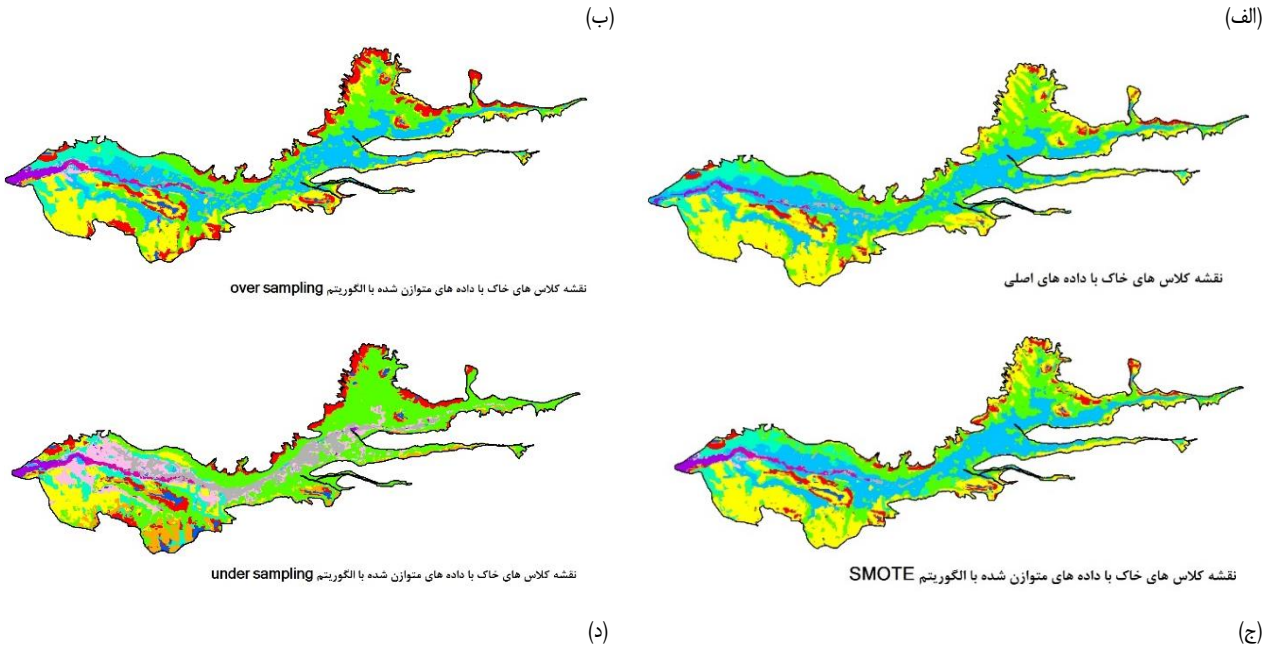
نقشه‌های پیش‌بینی‌شده کلاس‌های خاک با استفاده داده‌های اصلی و سه رویکرد استفاده شده جهت کاهش اثرات داده‌های نامتوازن و با استفاده از مدل جنگل تصادفی و مدل svm در شکل‌های ۴ تا ۷ آورده شده است.

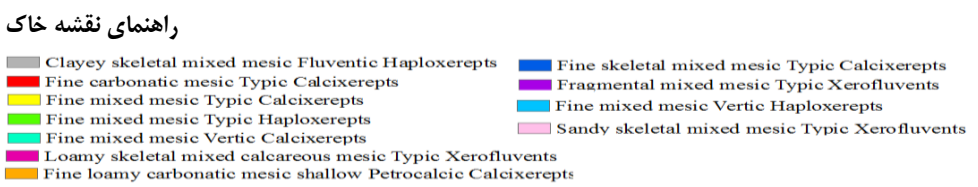


شکل ۴. نقشه رقومی کلاس‌های خاک با رویکردهای اول و دوم و استفاده از مدل RF

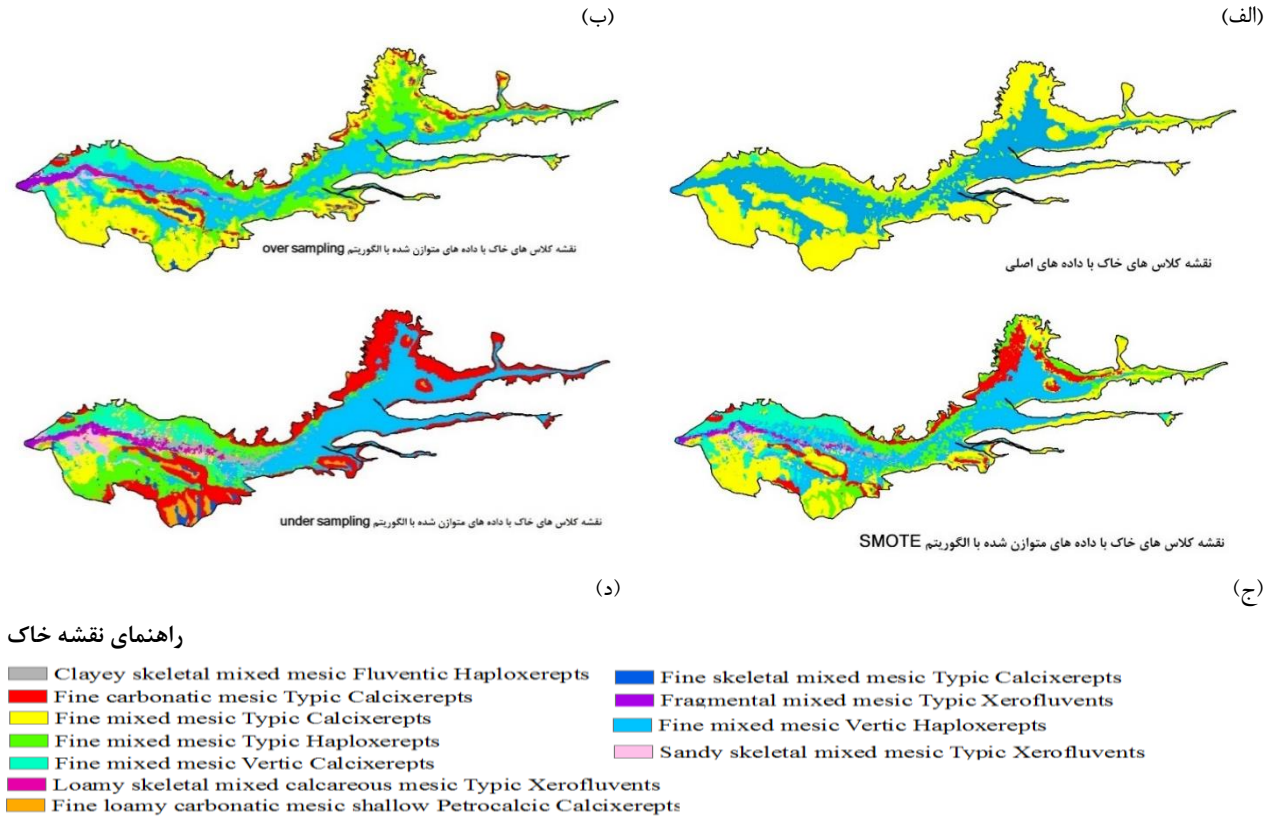


شکل ۵. نقشه رقومی کلاس‌های خاک با رویکردهای اول و دوم و استفاده از مدل SVM





شکل ۶. نقشه رقومی کلاس‌های خاک با داده‌های اصلی (الف)و الگوریتم‌های بازنمونه‌گیری (رویکرد سوم)(ب، ج، د) با استفاده از مدل RF



شکل ۷. نقشه رقومی کلاس‌های خاک با داده‌های اصلی (الف) و الگوریتم‌های بازنمونه‌گیری (رویکرد سوم) (ب، ج، د) با استفاده از مدل SVM

دقت پیش‌بینی‌های انجام‌شده توسط دو مدل برازش شده و رویکردهای موردبررسی با استفاده ماتریس خطای متشکل از مقادیر واقعی حاصل از ۳۵ نقطه مطالعاتی مشارکت داده نشده در آموزش مدل (داده‌های اعتبار سنجی) و مقادیر پیش‌بینی‌شده توسط مدل در جدول شماره ۴ ارائه‌شده است.

نتایج حاصل از مقایسه دو آماره صحت عمومی و شاخص کاپای مدل آموزشی و مدل اعتبار سنجی بیانگر بزرگ‌تر بودن این دو شاخص در مدل‌های آموزشی می‌باشد. علت این موضوع اینست که در اعتبارسنجی مدل آموزشی در بررسی و ارزیابی مدل تکیه‌بر داده‌هایی است که مشاهده‌شده و در ساختن مدل به‌کار رفته‌اند (۶۰ خاک‌رخ‌آموزشی)، اما اعتبارسنجی مدل پیش‌بینی بر اساس داده‌هایی است که مشاهده شده‌اند ولی در هنگام ساختن مدل به کار نرفته‌اند (۳۵ خاک‌رخ اعتبارسنجی). این داده‌ها به منظور بررسی و سنجش کارایی مدل برای پیش‌بینی داده‌های جدید به کار می‌روند. بنابراین، با توجه به نامتوازن بوده داده‌ها (کلاس‌های خاک) ازنظر فراوانی و توزیع در منطقه مطالعاتی و درنتیجه آن عدم قرار گرفتن برخی از کلاس‌ها در هر دودسته آموزش و اعتبارسنجی این نتیجه مورد انتظار می‌باشد.

در مقایسه عملکرد دو مدل مورداستفاده (RF و SVM) همان‌طور که در جدول ۴ ارائه شده است با توجه به شباهت نسبی بین مقادیر صحت عمومی و شاخص کاپا، می‌توان این‌گونه بیان کرد که این دو مدل از لحاظ آماره‌های صحت‌سنجی برتری کلی نسبت به هم نداشته‌اند. اما نقشه‌های رقومی تولید شده توسط دو مدل نشان می‌دهد که مدل SVM با توجه به ماهیت آن در استفاده از خطوط و صفحات برای طبقه‌بندی دادها در پیش‌بینی کلاس‌های خاک با فراوانی کم ناتوان بوده است. به‌طوری‌که در پیش‌بینی کلاس‌های خاک با داده‌های اصلی فقط قادر به پیش‌بینی سه کلاس A، B و C به ترتیب با فراوانی‌های ۱۷، ۱۴ و ۱۴ بوده است.

جدول ۴ . مقادیر صحت عمومی و کاپا دو مدل برازش شده در رویکردهای مورد بررسی

مدل	رویکرد مورد بررسی		مدل آموزشی		مدل پیش‌بینی	
			صحت عمومی	شاخص کاپا	صحت عمومی	شاخص کاپا
RF	داده‌های اصلی		۰/۵۲	۰/۳۶	۰/۲۸	۰/۱۵
	رویکرد اول		۰/۴۸	۰/۳۲	۰/۲۲	۰/۱۳
	رویکرد دوم		۰/۶۴	۰/۳۸	۰/۴۸	۰/۱۸
	رویکرد سوم	Over Sampling	۰/۳۴	۰/۱۸	۰/۲۸	۰/۱۷
		Under Sampling	۰/۲۱	۰/۰۶	۰/۱۴	۰/۰۴
		SMOTE	۰/۴۱	۰/۳۱	۰/۲۶	۰/۱۴
SVM	داده‌های اصلی		۰/۴۵	۰/۲۹	۰/۲۵	۰/۱۲
	رویکرد اول		۰/۵۰	۰/۳۳	۰/۳۱	۰/۱۴
	رویکرد دوم		۰/۶۲	۰/۳۱	۰/۵۲	۰/۲۲
	رویکرد سوم	Over Sampling	۰/۳۵	۰/۱۸	۰/۲۲	۰/۰۸
		Under Sampling	۰/۲۳	۰/۰۸	۰/۲۲	۰/۱۳
		SMOTE	۰/۴۲	۰/۲۲	۰/۳۱	۰/۲۱

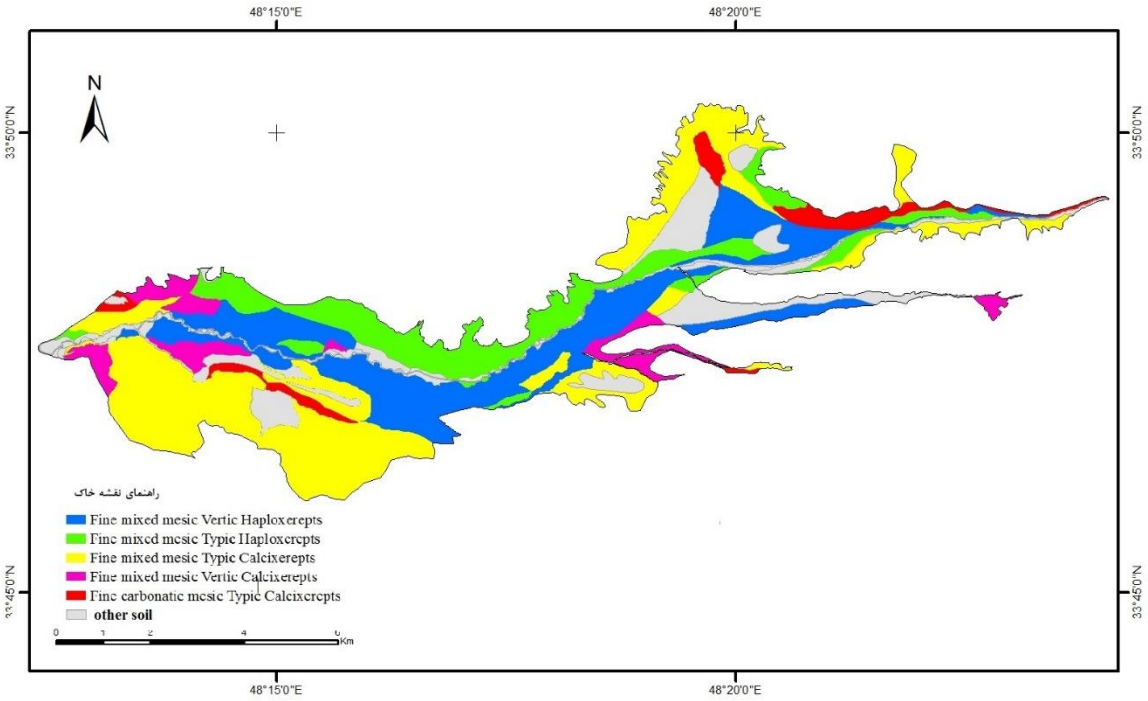
نتایج حاصل از مقایسه نقشه‌های تولیدشده توسط سه الگوریتم بازنمونه‌گیری برای بهبود کلاس‌بندی داده‌های نامتوازن در آموزش مدل‌ها (رویکرد سوم) نشان‌دهنده عملکرد ضعیف دو الگوریتم Over Sampling و Under Sampling می‌باشد. علت این موضوع حذف برخی از کلاس‌های خاک غالب منطقه مطالعاتی به علت کم نمونه‌گیری از کلاس‌های غالب (درروش Under Sampling) و اضافه شدن نمونه به کلاس‌های اقلیت (درروشOver Sampling) در نقشه‌های پیش‌بینی‌شده توسط مدل می‌باشد. علت دیگر این موضوع می‌تواند تصادفی بودن انتخاب نمونه‌ها برای آموزش مدل در این روش‌ها باشد. به عبارت دیگر در حذف کلاس‌های خاک غالب در منطقه مطالعاتی در کم‌نمونه‌گیری و یا برداشت بیشتر از نمونه‌های با فراوانی کم در بیش‌نمونه‌گیری، اهمیت خاک‌های غالب منطقه مطالعاتی و همچنین توزیع جغرافیایی آن‌ها در نظر گرفته نمی‌شود نمونه‌برداری ها بصورت کاملا تصادفی صورت می‌گیرد. درصورتی‌که در انجام نمونه‌برداری برای مطالعه خاکشناسی برداشت نمونه‌ها به‌صورت کاملاً تصادفی نیست و نمونه‌برداری‌ها با توجه به نقشه اولیه که در این مطالعه نقشه ژئوفرم‌های منطقه مطالعاتی می‌باشد، صورت گرفته است.

اما روش SMOTE به علت ایجاد نمونه‌های مصنوعی جدید از کلاس اقلیت و عدم حذف از نمونه‌های کلاس اکثریت نقشه‌های قابل‌قبول‌تری را ارائه کرده است و مقادیر صحت عمومی و شاخص کاپای مدل برازش شده به‌وسیله این الگوریتم نسبت به دو الگوریتم دیگر موید این موضوع می‌باشد. تقی‌زاده و همکاران (۲۰۱۵) در پژوهشی که باهدف استفاده از استراتژی‌های بازنمونه‌گیری و یادگیری ماشین در تهیه نقشه رقومی خاک‌های ایران، انجام دادند این نتیجه را تائید کرده‌اند.

نتایج حاصل از مقایسه دو رویکرد اول که در آن‌ها به طور دستی نسبت به طبقه‌بندی کلاس‌های خاک منطقه مطالعاتی برای رسیدن به کلاس‌های متوازن‌تر ازلحاظ فراوانی اقدام شده بود، نشان‌دهنده عملکرد بهتر مدل RF در رویکرد دوم (طبقه‌بندی کلاس‌های خاک ازلحاظ مدیریتی) می‌باشد. نکته قابل‌توجه این است که مدل SVM علیرغم بالا بودن شاخص‌های اعتبارسنجی هنگام استفاده از این رویکرد، نقشه‌ای فقط با دو کلاس دارای فراوانی بیشتر تولید کرده است. علت این موضوع همان‌طور که در قبل اشاره شد این است که SVM الگوریتمی است که در آن کلاس‌ها به‌وسیله یک ابرصفحه جداکننده که روی‌داده‌های آموزشی تعریف می‌شود، از هم جدا و

مشـــخص می‌شـــوند (Weston and Watkins, 1998). بنابراین به علت فراوانی کم نمونه‌های در کلاس‌های MO3 و MO4 الگوریتم SVM توانایی طبقه‌بندی درست همه نمونه‌ها را نداشته است.

نتایج تحقیق حاضـر نشـان داد که دقت الگوریتم تنها محرک انتخاب بهترین الگوریتم نیسـت و عدم قطعیت نقشـه‌های تولیدشـده از اهمیت زیادی برخوردار اسـت، که بایسـتی در مطالعات آتی موردتوجه قرار گیرد. در این مطالعه نقشـه‌های کلاس‌های تولیدشـده توسـط الگوریتم‌های یادگیری ماشـین به‌صـورت بصـری با نقشـه خاک تهیه‌شـده به روش مرسـوم برای منطقه مطالعاتی مقایسـه شـد (Eftekhari, 2018) (شکل ۸).



شکل ۷. نقشه کلاس‌های خاک منطقه مطالعاتی تهیه شده به روش مرسوم نقشه‌برداری خاک

همانطور که در شکل ۸ ارائه شده است علیرغم نتایج ضعیف حاصل از آماره‌های اعتبارسنجی، شباهت توزیع و گسترش خاک‌های کلاس‌های A، B و C که فراوانترین خاک‌های منطقه مطالعاتی بوده و به ترتیب در نقشـــه‌ها با رنگ‌های آبی، سـبز و زرد نشـان داده شده‌اند، در نقشـه‌های حاصـل از مدل RF توسـط مجموعه داده‌های اصـلی و همچنین الگوریتم بازنمونه‌گیری SMOTE، با نقشـه خاک تهیه‌شده به روش مرسوم منطقه مطالعاتی قابل‌توجه بود. بنابراین یکی از دلایل نتایج ضعیف اعتبارسنجی مدل‌ها می‌تواند مربوط به عدم توانایی مدل‌ها در پیش‌بینی سایر کلاس‌های خاک منطقه مطالعاتی باشد.

۴. بحث و نتیجه‌گیری

این مطالعه باهدف ارزیابی توانایی دو الگوریتم یادگیری ماشـین RF و SVM برای نقشـه‌برداری رقومی کلاس‌های خاک در سـطح فامیل انجام گرفت. همچنین، موضـوع عدم توازن تعداد مشـاهدات مربوط به کلاس‌های خاک با اسـتفاده از ۶ مجموعه داده، ازجمله مجموعه داده‌های اصـلی خاک و پنج مجموعه داده ایجادشـده توسـط چندین رویکرد نمونه‌گیری مجدد، روی داده‌های اصـلی قبل از مدل‌سـازی

موردبررسی قرار گرفت. در بین دو الگوریتم موردمطالعه، مدل جنگل تصادفی دقت بیشتری را با توجه به‌دقت کلی و مقدار کاپا نشان داد و همچنین موضوع عدم پیش بینی کلاس‌های خاک اقلیت در نقشه‌های تولیدشده توسط SVM اتفاق افتاده بود. به‌طورکلی می‌توان عنوان کرد که افزایش تعداد کلاس‌های خاک در سطح فامیل خاک و درنتیجه کاهش فراوانی آن‌ها سبب می‌شود آموزش مدل‌ها برای کلاس‌های با فراوانی کم به‌خوبی صورت نگیرد و این مسئله منجر به کاهش قابل‌توجه صحت کلی مدل‌ها می‌شود. اما نقشه رقومی تولیدشده حاکی از صحت قابل‌توجه مدل‌ها در پیش‌بینی پراکنش مکانی کلاس‌های با فراوانی مناسب است. مقایسه الگوریتم‌های بازنمونه‌گیری از مجموعه داده‌های اصلی حاکی از ناکارآمد بودن دو روش Over Sampling و Under Sampling در تهیه نقشه کلاس‌های خاک منطقه مطالعاتی بود. اما ترکیب روش نمونه‌برداری مجدد SMOTE و مدل جنگل تصادفی، نقشه خاک قابل قبولی را ارائه داد. این موضوع با مقایسه نقشه‌های رقومی حاصل شده با نقشه خاک منطقه مطالعاتی قابل تائید می‌باشد.

به‌طورکلی می‌توان استنباط کرد که میزان تنوع خاک‌ها در منطقه موردمطالعه، تراکم نمونه‌برداری و نوع پارامترهای محیطی مورداستفاده از مهم‌ترین عواملی هستند که صحت پیش‌بینی کلاس‌ها را تحت تأثیر قرار می‌دهند. بنابراین پیشنهاد می‌گردد در مطالعات پیش‌رو صحت پیش‌بینی‌ها برای هرکدام از کلاس‌های خاک به‌صورت جداگانه موردبررسی قرار گیرد. همچنین با توجه به اینکه نقشه حاصل از نقشه‌برداری رقومی خاک صرفاً نقشه‌رده‌بندی خاک می‌باشد، پیشنهاد می‌شود جهت ارائه نقشه خاک مدیریت‌پذیر منطقه مطالعاتی، ویژگی‌های اثرگذار بر مدیریت اراضی نیز موردتوجه قرار گیرند و درنهایت باید به این نکته توجه داشت که اگرچه توسعه روش‌های جدید نقشه‌برداری خاک ضروری است، اما دانش تخصصی و آگاهی از عوامل مؤثر بر تغییرپذیری مکانی خاک‌ها نقش بسیار مهمی در تهیه نقشه‌های خاک با دقت بالا دارد.

Reference

Abdi, L., & Hashemi, S. (2015). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE transactions on Knowledge and Data Engineering*, *28*(1), 238-251.

Adhikari, K., Minasny, B., Greve, M. B., & Greve, M. H. (2014). Constructing a soil class map of Denmark based on the FAO legend using digital techniques. *Geoderma*, *214*, 101-113.

Bagheri Bodaghabadi, M., Martínez-Casasnovas, J. A., Esfandiarpour Borujeni, I., Salehi, M. H., Mohammadi, J., & Toomanian, N. (2016). Database extension for digital soil mapping using artificial neural networks. *Arabian Journal of Geosciences*, 9(18), 1-13.

Banai M.H. (1998). Soil thermal-moisture regimes map of Iran on 1:1,250,000 scales. Soil and Water Research Institute, Tehran. Iran. (In Persian)

Brungard, C. W., J. L. Boettinger, M. C. Duniway, S. A. Wills & T. C. Edwards. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*. 240: 68–83.

Chawla, N. V., K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer. (2002). "SMOTE: Synthetic Minority Over-Sampling Technique". *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357.

Congalton, R. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing Environmental*, 37: 35–46.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Syst*, Appl.73, 220–239.

Hengl, T. (2007). A Practical Guide to Geostatistical Mapping of Environmental Variables. EUR 22904 EN. Luxembourg (Luxembourg): Office for Official Publications of the European Communities. JRC38153.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I.,Mantel, S., & Kempen, B.(2017). SoilGrids250m: global gridded soil information based on machine learning. *PLoS One*, 12, e0169748.

Eftekhari, k. (2018). Detailed soil survey and land classification of Honam sub catchment, emphasizing on soil-landform relations. Soil and water reserch institute. Project No: 14-10-10-9201-92001-K9201. (In Persian)

Eftekhari, k. (2019). Discriminating of Geomorphic Surfaces in Honam Sub catchment as a Basis for Delineating Homogenous Land Areas, Using Landscape Analysis Techniques. Soil and water research institute. Project No: 14-10-10-9201-92002-K9201. (In Persian)

Jackson ML. (1973). Soil chemical analysis. New Delhi: Prentice Hall of India Pvt. Ltd.

Jafari A., Ayoubi S.H., Khademi H., Finke P.K., & Toomanian N. (2013). Selection of a taxonomic level for soil mapping using diversity and map purity indices: A case study from an Iranian arid region. *Geomorphology*, 201: 86-97.

Jasiewicz, J., & Stepinski, T. F. (2013). Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182, 147-156.

Kovačević, M., Bajat, B., & Gajić, B. (2010). Soil type classification and estimation of soil properties using support vector machines. *Geoderma*, 154(3-4), 340-347.

Ma, T., Brus, D. J., Zhu, A. X., Zhang, L., & Scholten, T. (2020). Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. *Geoderma*, 370, 114366.

Martinez-Taboada, F., & Redondo, J. I. (2020). Variable importance plot (mean decrease accuracy and mean decrease Gini). *Plos One*, 15(4), e0230799.

Massawe, B. H., Subburayalu, S. K., Kaaya, A. K., Winowiecki, L., & Slater, B. K. (2018). Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. *Geoderma*, 311, 143-148.

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

Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, 264, 301-311.

Mirakzehi, M., Pahlavan-Rad, M. R., Shahriari, A, & Bameri, A. (2018). Digital soil mapping of deltaic soils: a case of study from Hirmand (Helmand) river delta, *Geoderma*, 313, 233–240.

Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E., & Mehnatkesh, A. (2016). The effectiveness of digital soil mapping to predict soil properties over low-relief areas. *Environmental monitoring and assessment*, 188(3), 1-13.

Mousavi, S.R., Sarmadian, F & Rahmani, A. (2020). Modelling and Prediction of Soil Classes Using Boosting Regression Tree and Random Forests Machine Learning Algorithms in Some Part of Qazvin Plain. *Iranian Journal of Soil and Water Research*, *50*(10), pp.2525-2538. (In Persian)

Nayal, A., Jomaa, H., & Awad, M. (2017). KerMinSVM for imbalanced datasets with a case study on arabic comics classification. *Engineering Applications of Artificial Intelligence*, 59, 159-169.

Nazari, S., Rostaminia, M., Ayoubi, S., Rahmani, A. & Mousavi, S.R. (2020). Efficiency of Different Feature Selection Methods in Digital Mapping of Subgroup and Soil Family Classes with Data Mining Algorithms. *Journal of Water and Soil*, 34(4,) 973-987. (In Persian)

Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A., & Sharififar, A. (2021). Digital mapping of soil classes using spatial extrapolation with imbalanced data. *Geoderma Regional*, 26, e00422.

pahlavan-Rad, M.R., Tomanian, N & Khormali, F (2016). Introduction to digital soil mapping. *Land Management Journal*, 4 (2), 97-114. (In Persian)

Pahlavan-Rad, M.R. & Akbari Moghaddam, A. R. (2018). Spatial variability of soil texture fractions and pH in a flood plain (a case study from eastern Iran), *Catena*, 160, 275–281.

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., & Thompson, J. (2018). Soil property and class maps of the conterminous United States at 100-meter spatial resolution. *Soil Science Society of America Journal*, *82*(1), 186-201.

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

Soil Survey Staff. (2014). Keys to Soil Taxonomy. 12th Edition, USDA-Natural Resources Conservation Service, Washington DC.

Taghizadeh-Mehrjardi, R., K. Nabiollahi, B. Minasny & J. Triantafilis. (2015). Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma*, 253-254: 67–77

Thomas, G. W. (1996). Soil pH and Soil Acidity. In: Methods of Soil Analysis, Part 3. Chemical Methods, American Society of Agronomy, Inc. Madison, Wisconsin, USA. 475-491

Weston, J., & Watkins, C. (1998). Multi-class support vector machines (pp. 98-04). Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May.

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.