Research Article

Agricultural Engineering., 46(2) (2023) 141-157
DOI: 10.22055/AGEN.2023.44173.1676

ISSN (E): 2588-526X
ISSN (P): 2588-5944

Evaluation of different feature selection algorithms for improving the spatial prediction of soil classes

V. Sadeghizadeh[1], S.A Abtahi[2,*], M. Baghernejad[3], A. Jafari[4] and S.A. Akbar Moosavi[5]

1. Ph.D. Student, Department of Soil Science, College of Agriculture, Shiraz University, Shiraz, Iran
2. Professor, Department of Soil Science, College of Agriculture, Shiraz University, Shiraz, Iran
3. Professor, Department of Soil Science, College of Agriculture, Shiraz University, Shiraz, Iran
4. Associate Professor, Department of Soil Science, College of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran
5. Professor, Department of Soil Science, College of Agriculture, Shiraz University, Shiraz, Iran

Received: 29 June 2023 Accepted: 12 August 2023 *Corresponding Author: seyedaliabtahi@yahoo.com

Abstract

Introduction: The number of environmental variables used in digital soil mapping has increased rapidly, which has made it a challenge to select and focus on the most important covariates. No environmental covariates have the same predictability in modeling, and some covariates may introduce noise that reduces the predictive power of the models used. On the other hand, it is beneficial to identify all environmental variables to obtain spatial information that can improve predictions. In this regard, the feature-selection- algorithms help reduce the dimensions of the predictive model by identifying the associated covariates. Therefore, this study aims to investigate different feature- selection- algorithms in the selection of auxiliary variables and evaluation their effect on the predictive model.

Materials and Methods: The area under study is a part of Darab city in the southeast of Fars province with an area of about 31000 hectares. In the study area 140 profiles were examined and excavated according to the diversity of geomorphological units and thus the type of soils. After excavating the profiles and checking the morphological characteristics of each soil profile, a sufficient amount of soil samples were collected from the genetic horizons and transported to the laboratory for further analysis. Some of the physical and chemical parameters of soils were tested using accepted techniques after air drying and passing through a 2 mm sieve. Finally, all profiles up to the great group level were classified using the most recent version of U.S. Soil Taxonomy system based on the data collected from field observations and the outcomes of laboratory analysis. Environmental variables include the parameters derived from the Digital Elevation Model, Landsat 8 images, geology, and geomorphology maps of the study area. All parameters were derived using ArcGIS (10.7), SAGAGIS (8.3) and ENVI (5.3) softwares. In the present study, four different feature selection techniques including Variance Inflation Factor (VIF), Principal Component Analysis (PCA), Boruta, and Recursive Feature Elimination (RFE), were used to identify an optimal set of covariates for predicting spatial classification of soil classes at the great group level. In addition, a Random Forest model (RF) with 10-fold cross-validation and the 5-repeat method, was used to compare different feature selection strategies in mapping soil classes.

The comparison of different feature selection techniques for estimating soil classes, was based on the evaluation criteria of accuracy and Kappa coefficient regarding observed and predicted classes.
Results and Discussion: The results showed that the prediction accuracy increased by using variables selected with different feature selection methods compared to using all variables in the model. In addition, the improvement in predictive performance is different between the four types of feature selection. The VIF and PCA methods had the highest and lowest accuracy index and Kappa coefficient, respectively. The Boruta method, with the lowest number of variables, improved the model's performance after the VIF method. However, the Kappa coefficient showed poor agreement between predicted and observed classes utilizing all approaches. The imbalance of soil classes could be a reason for decreasing the accuracy index and Kappa coefficient. However, the random forest model, with and without feature selection methods, identified all soil great groups in the study area. Therefore, it can be concluded that the Random Forest algorithm is a very powerful technique for spatial prediction of soil classes in the study area. Although the performance of the model varied using different feature selection algorithms, the predicted soil maps had similar spatial patterns. Based on the prediction of model with the variables selected by the VIF, the resulting map indicates that Ustorthents soils are mainly located in high altitude regions with steep slopes. Haplustepts, Calciustepts, and Calciusterts great groups have developed in places with low to medium slopes. Haplosalids have developed downstream of the salt dome. Great groups of Ustifluvents were discovered in fluvial sedimentary plains. Endoaquepts were found in the floodplains, which had the smallest area on the predicted map.
Conclusion: Overall, the findings indicate that the feature selection methods can utilize significant dependencies among relevant covariates to predict soil classes and to improve modeling accuracy. In the current study, the environmental factors, obtained from the Digital Elevation Model, were selected as key variables, showing the importance of topography and morphology in the classification of soil types in the area. Although the selected variables improved the performance of the model, the prediction of soil classes was random. This could be attributed to the imbalance of soil classes.

Keywords: Digital soil mapping, feature selection, covariates, random forest

مهندسی زراعی (مجله علمی کشاورزی) جلد ۴۶ شماره۲، تابستان ۱۴۰۲

مقاله پژوهشی

ارزیابی الگوریتم‌های مختلف انتخاب ویژگی در بهبود پیش‌بینی مکانی کلاس‌های خاک

وحیده صادقی زاده [۱]، سید علی ابطحی [*۲]، مجید باقر نژاد [۳]، اعظم جعفری [۴] و سید علی اکبر موسوی [۵]

۱- دانشجوی دکتری علوم خاک، دانشکده کشاورزی، دانشگاه شیراز، شیراز، ایران
۲- استاد گروه علوم خاک، دانشکده کشاورزی، دانشگاه شیراز ، شیراز، ایران
۳- استاد گروه علوم خاک، دانشکده کشاورزی، دانشگاه شیراز، شیراز، ایران
۴- دانشیار گروه علوم خاک، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، کرمان، ایران
۵- استاد گروه علوم خاک، دانشکده کشاورزی، دانشگاه شیراز، شیراز، ایران

تاریخچه مقاله

دریافت: ۱۴۰۲/۰۱/۰۸

پذیرش نهایی: ۱۴۰۲/۰۵/۲۱

کلمات کلیدی:
نقشه‌برداری رقومی خاک،
انتخاب ویژگی،
متغیرهای کمکی،
جنگل تصادفی

* عهده دار مکاتبات
Email: seyedaliabtahi@yahoo.com

چکیده

تعداد متغیرهای محیطی مورد استفاده برای نقشه‌برداری رقومی خاک به سرعت افزایش یافته است، که انتخاب و تمرکز بر روی مهم‌ترین متغیرهای کمکی را با چالش روبه‌رو کرده است. از طرفی، شناسایی همه متغیرهای محیطی به منظور دستیابی به اطلاعات مکانی برای بهبود پیش‌بینی‌ها، سودمند است. در این راستا، الگوریتم‌های انتخاب ویژگی با شناسایی متغیرهای کمکی مرتبط، به کاهش ابعاد مدل پیش‌بینی کننده کمک می‌کنند. در مطالعه حاضر، چهار تکنیک مختلف انتخاب ویژگی شامل عامل تورم واریانس (VIF)، تجزیه مولفه‌های اصلی (PCA)، باروتا (Boruta) و حذف ویژگی بازگشتی (RFE) به منظور تولید مجموعه‌ای بهینه از متغیرهای کمکی، برای پیش‌بینی مکانی کلاس‌های خاک در سطح گروه بزرگ به کمک مدل جنگل تصادفی بکار گرفته شد. مقایسه تکنیک‌های مختلف انتخاب ویژگی در تخمین کلاس‌های خاک، با استفاده از معیارهای ارزیابی دقت و ضریب کاپا بین مقادیر مشاهده‌شده و پیش‌بینی‌شده، انجام شد. نتایج نشان داد، با استفاده از متغیرهای انتخاب شده توسط روش‌های مختلف انتخاب ویژگی نسبت به کاربرد همه متغیرها در مدل، دقت پیش‌بینی تا حدودی افزایش یافت. همچنین در میان چهار رویکرد انتخاب ویژگی، بهبود عملکرد پیش‌بینی متفاوت بود. روش VIF و PCA به ترتیب بیشترین و کمترین دقت و ضریب کاپا را داشتند، در حالی که روش باروتا با کمترین تعداد متغیر توانست بعد از VIF عملکرد مدل را بهبود بخشد. به‌طور کلی یافته‌ها نشان داد، کاربرد روش‌های انتخاب ویژگی می‌تواند از وابستگی قابل‌توجه متغیرهای کمکی مربوطه برای پیش‌بینی کلاس‌های خاک استفاده کند و دقت مدل‌سازی را بهبود بخشد.

مقدمه

خاک یکی از ضروری‌ترین و محدودترین منابع زمین است و با فراهم کردن خدمات حیاتی اکوسیستم، از جمله تهیه غذا، تصفیه آب، کاهش آلاینده‌ها، چرخه مواد مغذی، ترسیب کربن، تنظیم آب و هوا و حفاظت از تنوع زیستی، زندگی بر روی زمین را قادر می‌سازد (۱۷، ۲۲). تحت فشار شدید رشد جمعیت، توسعه اقتصادی و تغییرات آب و هوایی، خاک‌های جهانی به‌طور مداوم تخریب می‌شوند. برای حفظ منابع خاک برای نسل بعدی، تقاضای فوری برای بهبود شیوه‌های مدیریتی وجود دارد که به اطلاعات مکانی صریح خاک برای تصمیم‌گیری نیاز دارد (۳۳). نقشه‌برداری خاک، طبقه‌بندی و مدل‌سازی پدولوژیک از اولین مطالعات علمی خاک‌ها بوده که سهم قابل‌توجهی در پیشرفت درک ما از خاک داشته است. با این حال، اکثر داده‌هایی که برای ایجاد نقشه‌های خاک استفاده می‌شوند، قدیمی هستند و بسیاری از آنها اطلاعات مربوط به نگرانی‌های اخیر جهانی را ارائه نمی‌دهند (۱). در ۲۰ سال گذشته، نقشه‌برداری خاک دستخوش یک انقلاب دیجیتالی شده است که ایده نقشه-برداری رقومی خاک[1] را به وجود آورد (۱۶).

مؤلفه های اصلی نقشه برداری رقومی "روش" و مجموعه ای از "متغیرهای محیطی" مورد استفاده برای پیش بینی کلاس ها و ویژگی های خاک است. جینی[2] (۱۲) پیشنهاد کرد که ماهیت و ویژگی¬های خاک در هر مکانی ناشی از تعامل پنج عامل تشکیل دهنده خاک، یعنی 'c'- اقلیم؛ 'o'- پوشش گیاهی و موجود زنده؛ 'r'- پستی و بلندی، توپوگرافی و ویژگی های زمین نما؛ 'p'- مواد مادری، سنگ شناسی؛ و 'a'- زمان یا سن است.

بر اساس عوامل تشکیل دهنده معادله ینی[3]، مک براتنی و همکاران[4] (۱۶) یک فرمول تجربی[5] برای تعیین رابطه کمی بین داده¬های مکانی و خاک پیشنهاد دادند. آنان هفت عامل یا متغیر محیطی را در مدل خود در نظر گرفتند که به عنوان "scorpan" شناخته می-شوند. پنج عامل تشکیل دهنده خاک از فرمول ینی هنوز هم به عنوان متغیرهای کمکی وجود دارد. دو متغیراضافی scorpan به طور خاص در مدل های پیش بینی مکانی قرار می گیرد، 's'- داده های صحرایی یا آزمایشگاهی خاک یا اطلاعات قبلی خاک در یک مکان؛ و 'n'- فضا، موقعیت مکانی یا نسبی است. اگر S کلاس خاک باشد، تابع f یک طبقه بندی نظارت شده یا برنامه یادگیری نظارت شده می¬باشد. برای پیش¬بینی کلاس خاک تابع f می تواند انواعی از روش¬ها باشد.

نقشه‌برداری رقومی خاک با ادغام بررسی‌های خاک، زمین‌آمار، سیستم اطلاعات جغرافیایی، سنجش از دور و یادگیری ماشین[6] به عنوان زیرشاخه‌ای در علم خاک در حال رشد سریع است (۱۸). مدل‌های خاک-زمین‌نما که خاک‌ها را تابعی از متغیرهای محیطی از جمله آب و هوا، موجودات زنده، توپوگرافی، مواد مادری و زمان توصیف می‌کنند، مبنای نظری برای نقشه‌برداری رقومی خاک هستند. عوامل خاص زمین‌نما شکل‌گیری ویژگی‌های خاک را تعیین می-کند. الگوهای توزیع مکانی انواع خاک یا ویژگی‌های آن ناشی از تأثیر متقابل عوامل سازنده خاک است (۱۶). این متغیرها می‌توانند عوامل سازنده خاک را از جنبه‌های مختلف مشخص کنند.

تعداد متغیرهای محیطی مورد استفاده برای نقشه‌برداری رقومی خاک به دلیل افزایش حجم داده‌های سنجش از دور، منابع متعدد مدل رقومی ارتفاع و داده‌های اقلیمی به سرعت افزایش یافته است، که انتخاب و تمرکز بر روی مهم‌ترین متغیرهای کمکی را با چالش روبه‌رو کرده است (۱). همه متغیرهای کمکی محیطی، قابلیت پیش‌بینی یکسانی در مدل‌سازی ندارند و برخی از متغیرهای کمکی ممکن است باعث ایجاد نویز شوند که توانایی پیش‌بینی مدل‌های مورد استفاده را کاهش می‌دهد (۲۸). با توجه به اینکه، جمع‌آوری همه متغیرهای محیطی به‌منظور دستیابی به اطلاعات مکانی

1- Digital Soil Mapping
2- Jenny
3- Yenny: S= f (c,o, r, p, t)
4- McBratney *et al.*
5- S = f (s, c, o, r, p, a, n)
6- Machine Learning

که می‌تواند برای پیش‌بینی‌ها ارزشمند باشد، سودمند است. الگوریتم‌های انتخاب ویژگی[1] با شناسایی متغیرهای کمکی مرتبط، به کاهش ابعاد ناشی از یک مجموعه متغیر کمکی بزرگ کمک می‌کنند. بنابراین، انتخاب متغیر عمدتاً قبل از برازش مدل پیش‌بینی انجام می‌شود. انتخاب ویژگی چندین مزیت دارد از جمله، ۱) واسنجی سریعتر مدل پیش‌بینی. ۲) کاهش پیچیدگی مدل. ۳) افزایش عملکرد مدل. ۴) اجتناب از چند خطی و (۵) تولید سریعتر نقشه (۴ و۲۹).

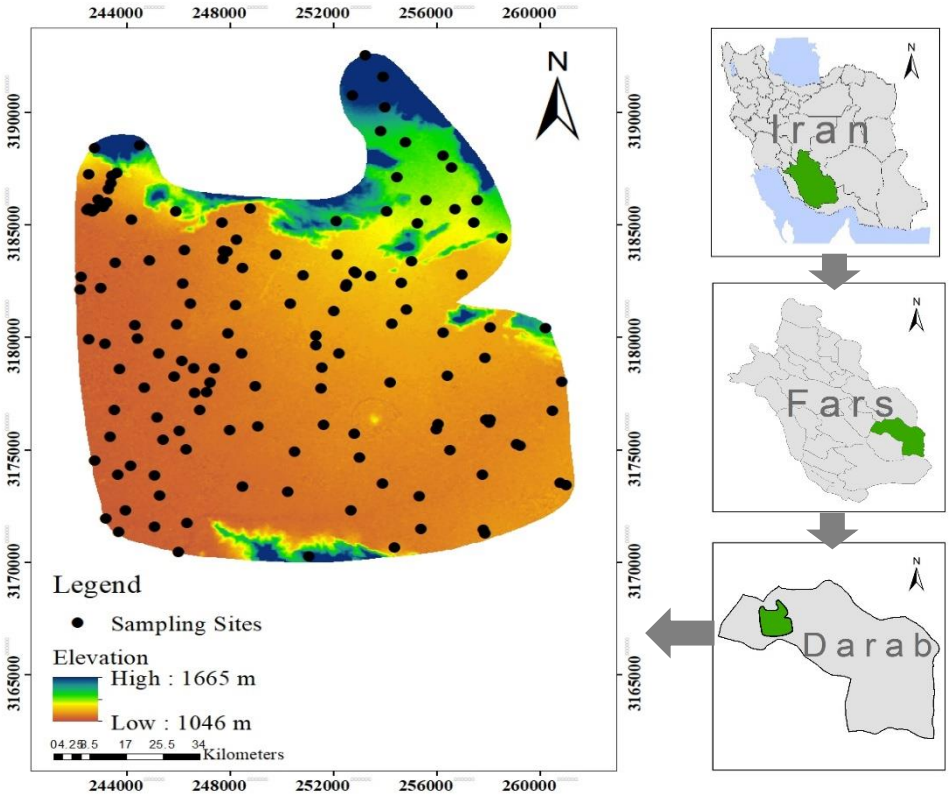
در حال حاضر، عمدتاً دو استراتژی برای انتخاب متغیر در نقشه‌برداری رقومی خاک وجود دارد، ۱) کاهش متغیرهای کمکی به عنوان یک مرحله پیش پردازش (قبل از واسنجی یک مدل یادگیری ماشین)، مانند انتخاب مرتبط‌ترین متغیرهای کمکی با ضریب همبستگی پیرسون بین خصوصیات خاک و متغیرهای کمکی، حذف متغیرهای کمکی که همبستگی بالایی با سایر متغیرهای کمکی دارند و یا حفظ چندین مؤلفه اول با استفاده از تحلیل مؤلفه‌های اصلی[2]. ۲) روش‌های بسته‌بند[3] که بر استنباط حاصل از کالیبره کردن یک مدل یادگیری ماشین برای ارزیابی اهمیت متغیر متکی هستند (۳۳). مطالعات نشان می‌دهد، استراتژی اول فقط رابطه خطی بین ویژگی‌های خاک و متغیرهای کمکی را در نظر می‌گیرد و یا حتی همبستگی‌های آنها را حذف می‌کند. بنابراین می‌تواند این متغیرهای کمکی را که به‌طور غیرخطی با ویژگی‌های خاک مرتبط هستند نادیده گیرد و در نتیجه عملکرد مدل را کاهش دهد (۲۹، ۳۱ و ۳۳). علاوه بر این، اگرچه مدل‌های یادگیری ماشین می‌توانند تحت‌تأثیر چند خطی بودن قرار بگیرند، اما همچنان قوی‌تر از رگرسیون خطی چندگانه سنتی هستند. بنابراین، استراتژی دوم برای مطالعات نقشه‌برداری رقومی خاک مبتنی بر یادگیری ماشین مناسب‌تر است و در میان این روش‌ها، محبوب‌ترین آنها باروتا[4] و حذف ویژگی بازگشتی[5] هستند (۳۳)، چرا که می‌توانند در شناسایی روابط پیچیده غیرخطی بین ویژگی خاک هدف و متغیرهای کمکی سودمند باشند (۸).

بنابراین، با توجه به اهمیت انتخاب متغیرهای کمکی پیش‌بینی‌کننده بر نتایج مدل‌سازی و تخمین، مطالعه حاضر با هدف بررسی الگوریتم‌های مختلف انتخاب ویژگی در انتخاب متغیرهای کمکی و ارزیابی تأثیر آن‌ها بر مدل پیش‌بینی‌کننده انجام شد.

مواد و روش‌ها

محدوده مورد مطالعه بخشی از شهرستان داراب در جنوب شرقی استان فارس به وسعت تقریباً ۳۱۰۰۰ هکتار، بین عرض جغرافیایی "۲۲ '۵۰ °۲۸ و "۱۲ '۳۷ °۲۸ شمالی و طول جغرافیای "۵۹ '۳۲ °۵۴ و "۳۷ '۲۱ °۵۴ شرقی است (شکل ۱). این منطقه دارای تابستان‌های گرم و طولانی و زمستان‌های نسبتاً معتدل و کوتاه، با میانگین بارندگی سالانه ۲۷۰ میلی‌متر و متوسط دمای سالانه حدوداً ۲۲ درجه سانتی‌گراد است، که براساس روش تعیین اقلیم گوسن جز مناطق نیمه بیابانی با رژیم‌های رطوبتی اریدیک-یوستیک و حرارتی هایپرترمیک می‌باشد. ساختار زمین‌شناسی منطقه شامل تشکیلات کرتاسه تا کواترنری بوده و از جمله سازندهای زمین‌شناسی آن، گنبد نمکی، سازند آهکی تابور، سازند دولومیتی جهرم، بخش‌های لغزشی سازند تابور، واحد چرت‌های رادیولیتی، سازند تبخیری ساچون، سازند آواری آغاجاری و سازند کنگلومرای بختیاری است (۱۰). براساس توپوگرافی و سنگ‌شناسی منطقه مورد نظر به واحدهای فیزیوگرافی واریزه‌های بادبزنی سنگریزه‌دار، مخروط افکنه‌های آبرفتی سنگریزه‌دار، دشت‌های دامنه‌ای و اراضی پست تقسیم شده است. پوشش گیاهی آن شامل گونه‌های مختلفی از جمله باغات مرکبات، زراعت پنبه، گندم، جو و انواعی از گیاهان مرتعی است.

1- Feature Selection
2- Principal Component Analysis
3- Wrapper
4- Boruta
5- Recursive Feature Elimination



شکل (۱) موقعیت منطقه مورد مطالعه و مکان‌های نمونه‌برداری

Figure (1) Location of the study area and sampling sites

روش نمونه‌برداری و تجزیه آزمایشگاهی

با توجه به نقشه ژئومورفولوژی، توپوگرافی و تفکیک واحدهای اراضی، تعیین نقاط نمونه‌برداری در محدوده مطالعاتی متناسب با تنوع واحدهای ژئومورفیک و در نتیجه تنوع خاک‌ها انجام شد. در نهایت، در هر یک از واحدهای مورد نظر باتوجه به الگوی نمونه¬برداری تصادفی، ۱۴۰ خاکرخ در طی فصول تابستان تا پاییز ۱۳۹۸-۱۳۹۹ حفر گردید. پس از حفر خاکرخ‌ها و بررسی ویژگی‌های مورفولوژیکی مربوط به هر خاکرخ (از قبیل ساختمان، رنگ و ...) از افق‌های ژنتیکی به مقدار کافی نمونه تهیه و بعد از هوا خشک شدن و عبور از الک ۲ میلیمتری، برخی از ویژگی‌های فیزیکی و شیمیایی خاک-ها از جمله، بافت خاک به روش هیدرومتر (۹)، ماده آلی به روش سوزاندن تر (۲۰)،کربنات کلسیم معادل به روش تیتراسیون برگشتی (۱۵)، اسیدیته خاک در گل اشباع توسط دستگاه pH متر و قابلیت هدایت الکتریکی در عصاره اشباع خاک توسط دستگاه هدایت سنج الکتریکی (۲۴)، ظرفیت تبادل کاتیونی (CEC) به روش استات سدیم (۲۷) تعیین شد. در نهایت براساس اطلاعات حاصل از مشاهدات صحرایی و نتایج تجزیه و تحلیل آزمایشگاهی، رده‌بندی تمام خاکرخ‌ها تا سطح گروه بزرگ بر اساس رده‌بندی آمریکایی نسخه ۲۰۱۴ تعیین گردید (۲۶).

متغیرهای محیطی

در مطالعه حاضر، ۲۴ متغیر محیطی از مدل رقومی ارتفاع[1]، تصاویر ماهواره لندست ۸، نقشه زمین‌شناسی و نقشه ژئومورفولوژی استخراج گردید (جدول ۱). از مدل رقومی ارتفاع تهیه شده از وبسایت مدل رقومی ارتفاع جهانی استر با وضوح مکانی ۳۰ متر، برای تولید ۹ متغیر محیطی مربوط به ویژگی‌های توپوگرافی منطقه استفاده شد. برخی شاخص‌های سنجش از دور شامل شاخص

1- Digital Elevation Model

پوشش گیاهی نرمال شده[1]، شاخص نسبت پوشش گیاهی[2]، شاخص تفاوت شوری نرمال شده[3] و ...، از تصاویر ماهواره ای لندست ۸ (مطابق با حداکثر زمان نمونه‌برداری) با اندازه پیکسل ۳۰ متر پردازش و استخراج گردید.

با زمین مرجع کردن مرزهای نقشه های زمین شناسی موجود از داراب و نمردان با مقیاس ۱:۱۰۰۰۰۰ (۱۰)، نقشه زمین شناسی منطقه مورد مطالعه رقومی شد. همچنین نقشه ژئومورفولوژی منطقه مطابق با روش جعفری و همکاران[4] (۱۱) تهیه گردید. این نقشه شامل سیمای اراضی، شکل اراضی، واحدهای سنگ شناسی و سطوح ژئومورفیک بود. در پایان وضوح مکانی همه متغیرها به ۳۰ متر تبدیل شد. در این مرحله، از نرم افزارهای ArcGIS (10.7)، SAGAGIS (8.3) و(5.3) ENVI استفاده شد.

انتخاب متغیرهای کمکی

قبل از شروع فرآیند مدل‌سازی، از ۴ الگوریتم انتخاب ویژگی به منظور حذف متغیرهای کمکی غیرضروری استفاده شد، این الگوریتم‌ها عبارتند از:

۱) تجزیه مولفه‌های اصلی (PCA): یک روش بدون نظارت برای انتخاب ویژگی است که به‌منظور کاهش ابعاد یک مجموعه داده و در عین حال تا حد امکان حفظ اطلاعات اصلی استفاده می‌شود (۲۱). برای اجرای این روش، از تابع prcomp() در بسته stats نرم افزار R برای محاسبه استفاده شد.

۲) باروتا (Boruta): یک روش انتخاب ویژگی و رتبه‌بندی ویژگی‌ها بر اساس الگوریتم جنگل تصادفی است و بنابراین می‌تواند روابط خطی و غیرخطی را شناسایی کند. هدف اصلی این استراتژی استفاده از تجزیه و تحلیل آماری و چندین اجرای جنگل تصادفی برای ارزیابی اهمیت متغیرهای پیش‌بینی‌کننده واقعی با متغیرهای تصادفی به اصطلاح سایه است. در هر اجرا، مجموعه متغیرهای پیش‌بینی با افزودن یک کپی از هر متغیر دو برابر می‌شود. در مجموعه داده‌های توسعه یافته، یک مدل جنگل تصادفی آموزش داده می‌شود و مقادیر اهمیت متغیر جمع‌آوری می‌شوند.

برای هر متغیر واقعی یک آزمون آماری انجام می‌شود تا اهمیت آن با حداکثر مقدار همه متغیرهای سایه مقایسه شود. متغیرهای با مقادیر اهمیت بزرگتر یا کوچکتر به ترتیب به عنوان مهم یا بکم اهمیت اعلام می‌شوند. پس از حذف همه متغیرهای غیرضروری و متغیرهای سایه، فرآیندهای قبلی تکرار می‌شوند تا زمانی که همه متغیرها دسته‌بندی شوند یا تعداد مشخصی از اجراها تکمیل شود (۶). برای تأیید انتخاب ویژگی، روش ما از بسته Boruta در زبان برنامه‌نویسی R پیروی کرده است.

۳) عامل تورم واریانس[5] (VIF): روشی است که معمولاً برای ارزیابی همخطی چندگانه بین پیش‌بینی کننده ها و انتخاب زیرمجموعه بهینه از پیش‌بینی کننده‌ها که همبستگی ندارند، استفاده می شود (۷). در فرآیند VIF، هر پیش‌بینی کننده با سایر پیش‌بینی‌های باقی‌مانده، از طریق یک مدل خطی با استفاده از رگرسیون کمترین مربعات معمولی، بررسی می‌شود. پس از محاسبه ضریب تعیین (R^2) از مدل خطی برازش شده، VIF با رابطه ۱ بدست می‌آید:

$$VIF_i = \frac{1}{1 - R_i^2} \qquad (۱)$$

1- Normalized Difference Vegetation Index
2- Ratio Vegetation Index
3- Normalized Difference Salinity Index
4- Jafari *et al.*
5 - Variance Inflation Factor

جدول(۱) متغیرهای محیطی مورد استفاده
Table(1) Environmental variables used

نام متغیر Variable Name	ماهیت Nature	فرمول Formula	اندازه Size	داده‌های محیطی Environmental Data
ارتفاع Elevation	توپوگرافی Topography		30	مدل رقومی ارتفاع DEM
آنالیز سایه Analytical Hillshading				
شکل اراضی Landforms				
شاخص تفکیک پایین دره MRVBF				
ارتفاع نرمال شده Normalized Height				
انحنای پروفیل Profile Curvature				
ارتفاع شیب Slope Height				
بافت Texture				
عمق دره Valley Depth				
شاخص روشنایی Bi	انعکاس نور Light Reflection	$BI = \sqrt{B_4^2 + B_5^2}$	30	تصویر لندست ۸ Landsat 8 Image
شاخص رنگ Ci		$CI = \frac{(B_4 - B3)}{(B_4 + B_3)}$		
شاخص تفاوت پوشش گیاهی Dvi		$DVI = B5 - B4$		
شاخص پایش محیطی جهانی Gemi		$\text{GEMI} = \frac{eta(1 - 0.25 * eta) - (B4 - 0.125)}{(1 - B4)}$ $\text{eta} = \frac{(2 * (B5^2 - B4^2) + 1.5 * B5 + 0.5 * B4)}{(B5 + B4 + 0.5)}$		
شاخص تفاوت کدورت نرمال شده Ndti		$NDTI = \frac{B4 - B3}{B4 + B3}$		
شاخص تفاوت پوشش گیاهی نرمال شده Ndvi		$NDVI = \frac{B5 - B4}{B + B4}$		

شاخص نسبت پوشش گیاهی Rvi	$RVI = \frac{B5}{B4}$	
شاخص گیاهی تعدیل شده خاک Savi	$SAVI = \frac{B5 - B4}{B5 + B4 + 0.5} * (1 + 0.5)$	
شاخص تفاوت آب نرمال شده Ndwi	$NDWI = \frac{B5 - B6}{B5 + B6}$	
شاخص کربنات CARBI	$CARBI = \frac{B4}{B3}$	
شاخص گچ GYPSI	$GYPSI = \frac{B6 - B5}{B6 + B5}$	
شاخص رس نرمال شده NDCI	$NDCI = \frac{B7 - B6}{B7 + B6}$	
شاخص تفاوت شوری نرمال شده NDSI	$NDSI = \frac{B4 - B5}{B4 + B5}$	
سنگ‌شناسی Lithology	نقشه پلیگونی Polygon Map	
ژئومورفولوژی Geomorphology	نقشه پلیگونی Polygon Map	
نقشه زمین شناسی Geology Map	1:100000	
نقشه ژئومورفولوژی Geomorphology Map	1:100000	

در ستون مربوط به فرمول، باندهای۳، ۴، ۵، ۶ و ۷ به ترتیب مربوط به باندهای سبز، قرمز،مادون قرمز نزدیک، مادون قرمز میانه و مادون قرمز دور تصویر لندست ۸ است.

In the formula column, Bands 3, 4, 5, 6 and 7 correspond to the green, red, near-infrared, middle-infrared, and far-infrared bands of the Landsat 8 image, respectively.

در اینجا R^2_i، R^2 برای پیش بینی i است. پیش‌بینی‌کننده‌های با مقادیر VIF بیش از ۱۰، با سایر پیش‌بینی‌کننده‌ها همبستگی بالایی دارند و بنابراین از مدل نهایی حذف می‌شوند (۳۳).از بسته usdm در R برای اجرای این الگوریتم استفاده شد.

۴) حذف ویژگی بازگشتی (RFE): از تمام ویژگی‌ها برای ساخت یک مدل استفاده می‌کند. در مرحله بعد، همکاری هر ویژگی در مدل را در فهرست ویژگی‌های رتبه بندی شده رتبه‌بندی می کند. RFE در نهایت ویژگی‌های نامرتبط که سهم بی‌معنی در مدل دارند را حذف می‌کند. علاوه بر این، RFE یک الگوریتم قدرتمند برای انتخاب ویژگی است که به مدل یادگیری خاص بستگی دارد (۳۰). اجرای این فرآیند در بسته Caret از زبان R صورت پذیرفت.

روش مدل سازی

مدل جنگل تصادفی[1] از مدل‌های پرکاربرد در داده-کاوی برای نقشه‌برداری رقومی خاک‌ها است. در مقایسه با سایر مدل‌های رگرسیون مانند درخت‌های تصمیم[2]، رگرسیون چندگانه[3] و ماشین‌های بردار پشتیبان[4]، جنگل تصادفی حساسیت کمتری نسبت به نویز داده‌ها داشته و قابلیت‌های یادگیری بالاتری دارد (۳۴). جنگل تصادفی یک الگوریتم طبقه‌بندی و رگرسیون با چندین درخت

1- Random Forest
2- Decision Trees
3- Multiple Regression
4- Support Vector Machines

است،که می‌تواند برای پیش‌بینی متغیر هدف در مکان-هایی که ناشناخته است، بر اساس روابط قبلاً تعریف شده بین متغیر هدف و پیش‌بینی کننده‌ها، مورد استفاده قرار گیرد. مروری از عملکرد مدل توسط بولستیکس و همکاران[1] (۲) ارائه شده است.

در این مطالعه، مدل جنگل تصادفی نه تنها به عنوان تابع انتخاب‌گر برای انتخاب متغیرهای بهینه شده، بلکه برای نقشه‌برداری کلاس خاک در سطح گروه بزرگ نیز استفاده شد. برای اجرای این مدل از بسته caret در نرم افزار R استفاده شد.

ارزیابی دقت پیش‌بینی

به منظور ارزیابی چهار نوع مختلف انتخاب ویژگی برای نقشه‌برداری در سطح گروه بزرگ خاک، متغیرهای انتخاب شده با این روش‌ها، برای آموزش مدل جنگل تصادفی مورد استفاده قرار گرفتند. دقت پیش‌بینی مدل با استفاده از اعتبارسنجی متقابل ده برابری[2] با ۵ تکرار ارزیابی شد (۲۳).

در این روش مجموعه داده‌های اندازه گیری شده به طور تصادفی به ۱۰ زیرمجموعه مساوی یا تقریباً هم اندازه تقسیم می‌شود، ۹ زیر مجموعه به عنوان مجموعه آموزشی انتخاب و بقیه به عنوان مجموعه آزمایش استفاده می‌شوند. اعتبار سنجی متقابل یک ساختار مدل‌سازی فراهم می‌کند که برای تقسیم چندین مجموعه داده آموزش و اعتبارسنجی استفاده می‌شود و تضمین می‌کند هر نمونه می‌تواند حداقل یک بار برای

اعتبار سنجی اختصاص یابد. بزرگترین مزیت این روش این است که به طور قابل اعتماد اجرا می‌شود و برای یک مجموعه نمونه کوچک بی‌طرف است (۳). بر این اساس، ۵۰ مجموعه آموزشی و ۵۰ مجموعه معتبر توسط داده‌های نمونه ایجاد شده که به ترتیب ۵۰ مدل و ۵۰ پیش‌بینی را به‌دست می‌آورند. میانگین دقت این ۵۰ پیش‌بینی برای ارزیابی روش‌های مختلف انتخاب ویژگی استفاده شد. دو شاخص آماری ضریب کاپا[3] و دقت کلی[4] که از رایج-ترین معیارهای مورد استفاده برای ارزیابی صحت طبقه-بندی هستند (۵)، برای ارزیابی مدل‌ها و دقت کاربر[5] و دقت تولید کننده[6] برای پیش‌بینی‌های هر کلاس خاک در سطح گروه بزرگ محاسبه شدند. این فرآیند را در بسته caret در نرم افزار R بهینه کردیم. شکل۲، مجموعه‌ای از پردازش‌های مورد استفاده در این مطالعه را نشان می‌دهد.

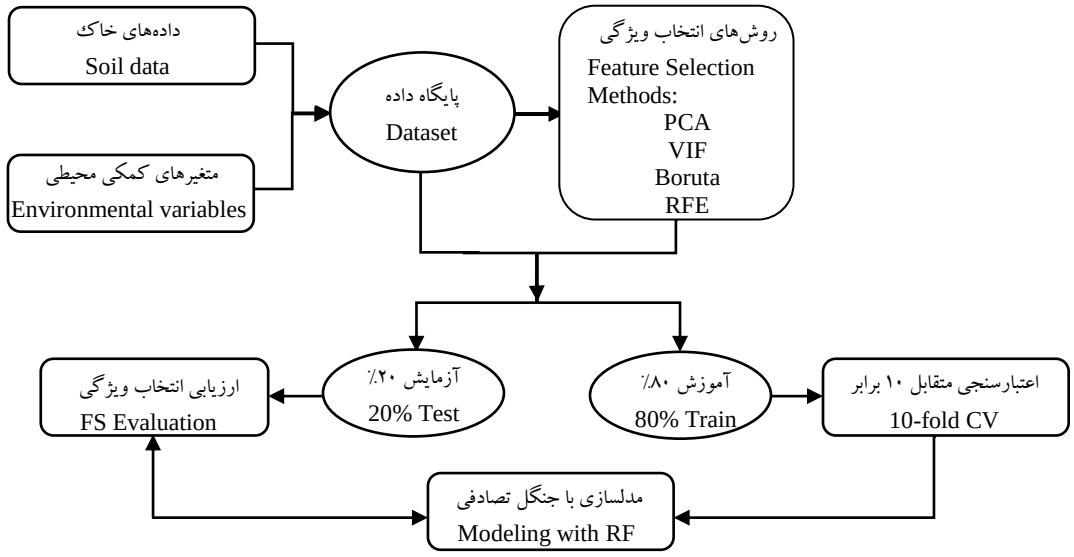
نتایج و بحث

انتخاب متغیرهای کمکی مهم برای نقشه‌برداری کلاس‌های خاک

وجود تفاوت‌های قابل توجه در تعداد و اهمیت متغیرهای انتخاب شده توسط چهار روش انتخاب ویژگی، برجسته است (شکل ۳). PCA بیشترین (۲۲ متغیر) و Boruta کمترین (۱۰ متغیر) تعداد متغیر را انتخاب کردند.

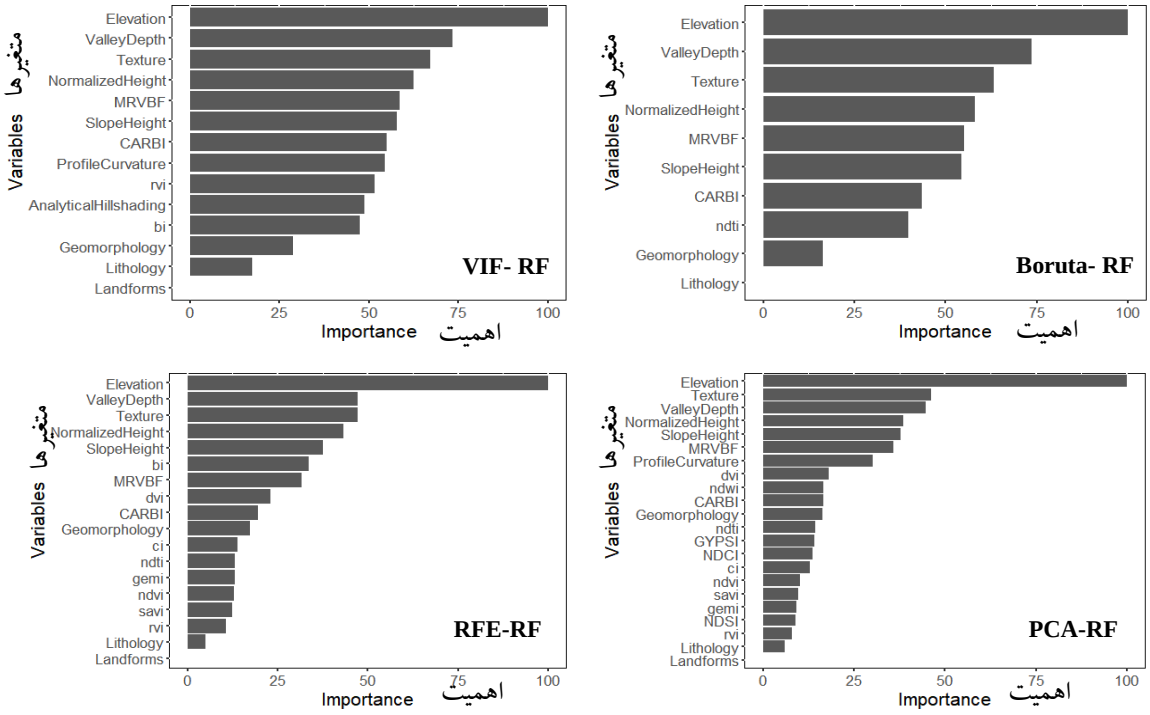
مجموعه متغیرهایی که توسط روش‌های مختلف انتخاب ویژگی استخراج گردید، متفاوت است، اما متغیرهای مهم انتخاب شده با یکدیگر سازگاری نسبی دارند. ۹ متغیر ارتفاع، شاخص تفکیک پایین دره، عمق دره، بافت، ارتفاع نرمال شده، ارتفاع شیب، شاخص کربنات، نقشه ژئومورفولوژی و نقشه زمین‌شناسی در مجموعه متغیرهای مهم انتخاب شده از روش‌های انتخاب ویژگی مشترک هستند. در میان این متغیرها، ۶ متغیر اول که مربوط به پارامترهای حاصل از مدل رقومی ارتفاع هستند، در همه روش‌ها از اهمیت بالاتری برخوردارند. برخی مطالعات پارامترهای استخراجی از مدل رقومی ارتفاع را به‌عنوان متغیرهای محیطی مناسب در مدل‌سازی کلاس‌ها و خصوصیات خاک بیان کردند (۱۱، ۱۳، ۳۲).

1- Boulesteix *et al.*
2- 10-Fold Cross-Validation
3- Kappa Coefficient
4- Overall Accuracy
5- User,s Accuracy
6- producer,s Accuracy



شکل(۲) فلوچارت روش‌شناختی، مجموعه‌ای از چارچوب‌های پردازش مورد استفاده در این مطالعه را توضیح می‌دهد.

Figure(2) Methodological flowchart explaining a series of processing frameworks employed for this study.



شکل(۳) اهمیت متغیرهای انتخاب شده توسط مدل جنگل تصادفی با روش‌های مختلف انتخاب ویژگی: (PCA-RF) و تجزیه مولفه‌های اصلی (RFE-RF)، حذف ویژگی بازگشتی (Boruta-RF)، باروتا (VIF-RF) عامل تورم واریانس

Figure(3) Importance of variables selected by the random forest model with different feature selection methods: VIF-RF, Boruta-RF, RFE-RF and PCA-RF

ارزیابی پیش‌بینی‌های مدل

نتایج جدول ۲ نشان می‌دهد، مدل جنگل تصادفی همراه با مجموعه متغیرهای مهم انتخاب شده توسط چهار الگوریتم انتخاب ویژگی، در مقایسه با مدل با مجموعه کامل متغیرها دقت و عملکرد نسبتاً بالاتری را ارائه می‌دهد. پیش‌بینی مدل با متغیرهای انتخاب شده به روش VIF و PCA به ترتیب بیشترین و کمترین دقت و ضریب کاپا را دارد، در حالی که روش باروتا با کمترین تعداد متغیر توانست بعد از VIF عملکرد مدل را افزایش دهد.

بهبود زیادی در دقت پیش‌بینی توسط متغیرهای انتخاب شده به روش VIF نسبت به استفاده از همه متغیرها در مدل مشاهده گردید، به‌طوری که میانگین دقت و ضریب کاپا به ترتیب ۲۳/۹۱٪ و ۵۰٪ افزایش یافت. پس از آن، افزایش دقت مشابهی در روش‌های انتخاب ویژگی Boruta و RFE رخ داده است. در پژوهشی، الگوریتم‌های جنگل تصادفی و رگرسیونی درختی توسعه‌یافته به همراه دو روش انتخاب متغیر VIF و PCA برای نقشه‌برداری کلاس خاک مورد استفاده قرار گرفت. نتایج نشان داد که استفاده از جنگل تصادفی و متغیرهای انتخاب شده از طریق VIF، باعث بهبود قابل توجهی در دقت نقشه‌برداری کلاس‌های خاک شده است (۱۹)، که با نتایج این مطالعه مطابقت دارد. ژانگ و همکاران[1] (۳۳) در یک مطالعه به بررسی ۴ الگوریتم مختلف برای انتخاب ویژگی پرداختند و بیان کردند که VIF کمترین عملکرد مدل را ارائه داده است. آن‌ها این نتیجه را ناشی از این دانستند که VIF تنها همبستگی بین متغیرهای پیش‌بینی‌کننده را محاسبه می‌کند و از ارتباط آن‌ها با متغیرهدف غافل می‌شود.

قابل ذکر است، در مطالعه حاضر مقادیر ضریب کاپا در همه روش‌ها به جز VIF، بیانگر توافق کم بین مقادیر پیش‌بینی و مشاهده شده است. به‌عبارتی، کلاس-های خاک توسط مدل به‌صورت تصادفی پیش‌بینی شده-اند که باعث بیش‌برآورد برخی کلاس‌های خاک و نادیده گرفتن برخی دیگر می‌شود. مقادیر کاپای بیش از ۰/۷۵ بیانگر توافق بالا یا پیش‌بینی منطقی، مقادیر بین۰/۴ تا ۰/۷۵ نشان‌گر توافق متوسط و مقادیر کمتر از ۰/۴ توافق کم یا تصادفی را نشان می‌دهند (۱۴). از دلایل کاهش دقت و ضریب کاپا را می‌توان به عدم تعادل کلاس‌های خاک در این پژوهش نسبت داد. شریفی‌فر و همکاران[2] (۲۵)، در مطالعه‌ای به بررسی نقشه‌برداری کلاس‌های نامتعادل خاک پرداختند، نتایج آن‌ها نشان داد، استفاده از داده‌های اصلی که دارای کلاس‌های نامتعادل برای نقشه‌برداری است، منجر به از دست رفتن کلاس اقلیت و مقادیر نسبتاً پایین ضریب کاپا برای برخی از مدل‌ها شده است.

الگوهای مکانی در نقشه‌های گروه بزرگ خاک

نقشه‌هایی که توسط مدل جنگل تصادفی با و بدون استفاده از روش‌های انتخاب ویژگی تولید شده‌اند، توانستند گروه‌های بزرگ خاک را در منطقه مورد مطالعه (شکل ۴) شناسایی کنند. از این رو، می‌توان نتیجه گرفت که الگوریتم جنگل تصادفی یک روش بسیار مفید برای پیش‌بینی مکانی کلاس‌های خاک در منطقه مورد مطالعه است.

با وجود تفاوت‌های مشاهده شده در عملکرد مدل با کاربرد روش‌های مختلف انتخاب ویژگی، نقشه‌های خاک پیش‌بینی شده، الگوهای مکانی مشابهی دارند و توزیع ناپیوسته‌ای را در مقیاس تحقیق نشان می‌دهند. این توزیع ناپیوسته با توپوگرافی منطقه به طور قابل‌توجهی هماهنگ است. همان‌گونه که قبلاً ذکر شد، مهمترین متغیرهای محیطی با استفاده از روش‌های انتخاب ویژگی، شامل ارتفاع، شاخص تفکیک پایین دره، عمق دره، بافت، ارتفاع نرمال شده و ارتفاع شیب هستند، که نشان‌دهنده این است توپوگرافی و مورفولوژی منطقه تأثیر بسزایی در متمایز کردن کلاس‌های خاک دارد (۱۱، ۱۳، ۳۲).

با توجه به اینکه مدل با متغیرهای انتخاب شده به روش VIF، دقت بیشتری در پیش‌بینی گروه بزرگ خاک

1- Zhang *et al.*

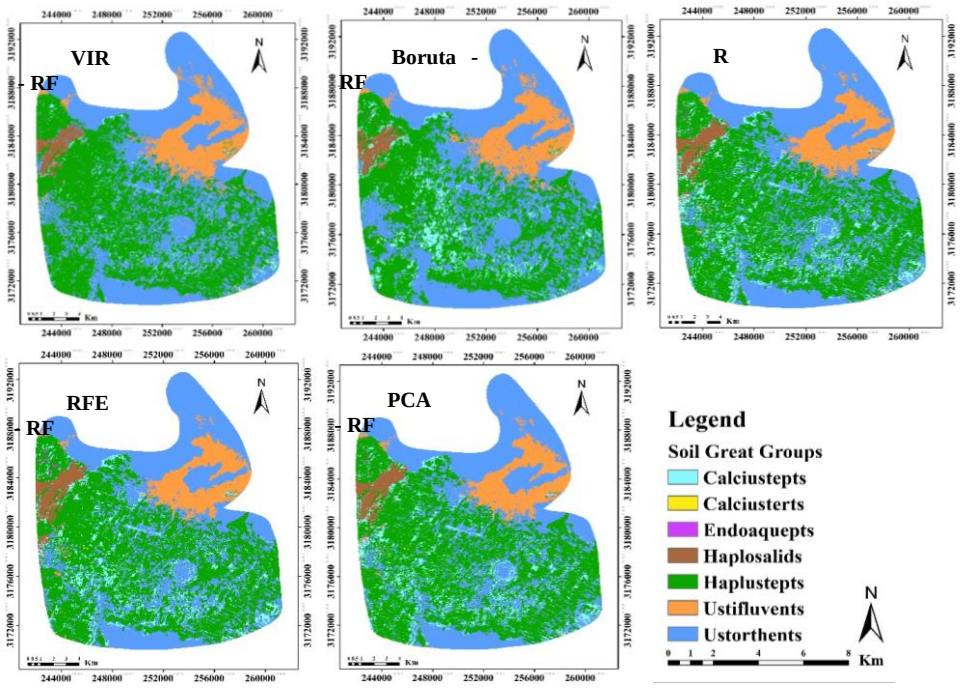
2 - Sharififar *et al.*

در منطقه مورد مطالعه دارد، جدول ۲، کلاس‌های خاک در سطح گروه بزرگ منطقه مورد مطالعه و نقشه حاصل از آن در شکل ۴، نشان می‌دهد خاک‌های یوست اورتنتز با دقت کاربر و دقت تولید کننده به ترتیب ۱ و ۰/۵، در مناطق مرتفع با شیب تند قرارگرفته‌اند. این خاک‌ها وسعت تقریباً زیادی از منطقه را در برگرفته و مستعد فرسایش-پذیری هستند و به همین دلیل جوان و تکامل کمی دارند. از طرف دیگر، گروه بزرگ‌های هاپلیوستپتز (که حداکثر وسعت منطقه را شامل می‌شود)، کلسی‌یوستپتز و کلسی-یوسترتز در مناطق با شیب کم تا متوسط تشکیل شده‌اند. در این مناطق به دلیل کشت و کار فراوان، امکان آبشویی کربنات‌ها در نیمرخ خاک وجود دارد و درنتیجه منجر به تشکیل این گروه از خاک‌ها با تکامل متوسط به بالا شده است.

در پایین دست گنبد نمکی، هاپلوسالیدزها توسعه یافته‌اند. به‌دلیل اینکه از نظر ژئومورفولوژی در پست‌ترین قسمت منطقه واقع شده‌اند، املاح شور توسط جریان‌های آبی وارد این قسمت از دشت می‌شوند و افق سالیک تشکیل شده است. این اراضی شوری خیلی زیادی دارند و بیشتر بصورت بایر و بیابانی می‌باشند. با وجود وسعت کم این خاک‌ها، مدل با دقت بسیار خوبی (هر دو دقت کاربر و دقت تولید کننده، ۱) توانسته هاپلوسالیدزها را پیش‌بینی کند. در شمال منطقه گروه بزرگ یوستی‌فلوونتز مشاهده می‌شود. این خاک‌ها در دشت‌های رسوبی رودخانه‌ای با مواد مادری آبرفتی و رسوبات قدیمی واقع شده است. در دشت‌های سیلابی بدلیل بالا بودن سطح آب زیرزمینی و ایجاد شرایط اکسید و احیایی خاک‌های اندواکوئپتز تشکیل شده‌اند، که کمترین مساحت نقشه پیش‌بینی (۰/۶۲ هکتار) متعلق به این گروه بزرگ خاک است. هر چند مدل انتخابی، دقت مناسبی در تشخیص خاک‌های کلسی-یوسترتز و اندواکوئپتز بدلیل تعداد مشاهدات میدانی خیلی کم این نوع خاک نسبت به سایر خاک‌ها نداشته است، اما به‌طور کلی می‌توان بیان کرد، نقشه حاصل از مدل، توزیع گروه‌های بزرگ خاک منطقه را به خوبی نشان می‌دهد و با مشاهدات میدانی همخوانی دارد. بنابراین انتخاب متغیر مناسب و توانایی مدل در ایجاد ارتباط بین متغیرهای کمکی محیطی و کلاس‌های خاک در پیش‌بینی‌ها مهم است.

جدول(۲) مقایسه پیش‌بینی‌های مدل با و بدون استفاده از روش‌های انتخاب ویژگی
Table(2) Comparison of model predictions with and without using feature selection methods

مدل پیش بینی Prediction Model	تعداد متغیر انتخاب شده Number Of Selected Variable	دقت کلی Overall Accuracy	ضریب کاپا Kappa Coefficient	تعداد متغیر انتخاب شده در هر گره درخت Mtry	تعداد درخت Ntree	خطای خارج از کیسه % Out of bag (OOB%)
RF	24	0.46	0.30	2	500	54.46
RF – PCA	22	0.47	0.30	2	500	53.68
RF – Boruta	10	0.50	0.36	2	500	46.43
RF – RFE	18	0.50	0.35	18	500	52.68
RF – VIF	14	0.57	0.45	2	500	50



شکل (٤) نقشه‌های پیش‌بینی گروه‌های بزرگ خاک توسط مدل جنگل تصادفی با و بدون استفاده از روش‌های انتخاب ویژگی

Figure (4) prediction maps of soil great groups by RF model with and

جدول(۳) کلاس‌های خاک در سطح گروه بزرگ منطقه مورد مطالعه براساس مدل RF - VIF

Table(3) soil classes in the great group level of study area based on the RF-VIF model

گروه بزرگ خاک Soli Great Group	مساحت Area (%)	مساحت Area (ha)	دقت کاربر User's Accuracy	دقت تولیدکننده producer's Accuracy
کلسی‌یوستپتز Calciustepts	2.516	779.935	0.5	0.2
کلسی‌یوسترتز Calciusterts	0.006	1.86	NA	NA
اندواکوئپتز Endoaquepts	0.002	0.62	NA	NA
هاپلوسالیدز Haplosalids	1.226	380.048	1	1
هاپلیوستپتز Haplustepts	51.4	15933.49	0.37	1
یوستی‌فلوونتز Ustifluvents	9.655	2992.953	0.8	0.66
یوست‌اورتنتز Ustorthents	35.195	10910.1	1	0.5
	100	30999		

نتیجه‌گیری

این پژوهش به بررسی برخی از الگوریتم‌های انتخاب ویژگی برای نقشه‌برداران خاک پرداخته است. در این پژوهش، چهار رویکرد مختلف انتخاب ویژگی برای پیش‌بینی گروه‌های بزرگ خاک با استفاده از مدل جنگل تصادفی مورد بررسی قرار گرفت. یافته‌ها نشان داد استفاده از روش‌های انتخاب ویژگی، می‌تواند از وابستگی قابل‌توجه متغیرهای کمکی مربوطه برای پیش‌بینی کلاس-های خاک استفاده کند و دقت مدل‌سازی را بهبود بخشد. در این روش‌ها، متغیرهای محیطی مشتق شده از مدل رقومی ارتفاع، به عنوان متغیر مهم انتخاب شدند، که نشان دهنده اهمیت توپوگرافی و مورفولوژی در تفکیک کلاس‌های خاک این منطقه بود. همچنین نقشه‌های حاصل از پیش‌بینی مدل، توزیع مکانی کلاس‌های خاک منطقه را به خوبی نشان دادند. هرچند استفاده از متغیرهای انتخاب شده توسط روش‌های مختلف انتخاب ویژگی نسبت به کاربرد همه متغیرها در مدل، عملکرد مدل را بهبود بخشید، اما به‌دلیل ضریب کاپای با تواق کم، پیش-بینی کلاس‌های خاک بصورت تصادفی بوده که احتمالاً به علت عدم تعادل کلاس‌های خاک است. بنابراین با توجه به نتایج حاصل از این پژوهش پیشنهاد می‌شود برای بهبود در مدلسازی و پیش‌بینی‌های منطقی، علاوه بر کاربرد الگوریتم‌های جدید انتخاب ویژگی، تکنیک‌های جایگزین دیگری مانند استفاده از دیگر معیارهای ارزیابی مدل، شناسایی و کاربرد سایر متغیرهای محیطی مؤثر بر تشکیل و تکامل خاک‌های منطقه، افزایش نقاط نمونه-برداری و روش‌های متوازن کردن داده‌ها اتخاذ گردد.

References

1. Arrouays, D., Lagacherie, P., and Hartemink, A. E. 2017. Digital soil mapping across the globe. Geoderma Regional, 9, 1-4.

2. Boulesteix, A. L., Janitza, S., Kruppa, J., and König, I. R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493-507.

3. Bouslihim, Y., Rochdi, A., and Paaza, N. E. A. 2021. Machine learning approaches for the prediction of soil aggregate stability. Heliyon, 7(3), e06480.

4. Chen, Y., Ma, L., Yu, D., Zhang, H., Feng, K., Wang, X., and Song, J. 2022. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. Ecological Indicators, 135, 108545.

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

6. Degenhardt, F., Seifert, S., and Szymczak, S. 2019. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics, 20(2), 492-503.

7. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... and Lautenbach, S. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46.

8. Ferhatoglu, C., and Miller, B. A. 2022. Choosing feature selection methods for spatial modeling of soil fertility properties at the field scale. In Proceedings of the 30th International Conference on Advances in Geographic Information Systems (pp. 1-2).

9. Gee G.W. and Bauder J.W. 1986. Particle size analysis. In: Klute A. Methods of Soil Analysis. Part 1. Physical properties. American Society of Agronomy. Madison. Wisconsin. pp: 383-411.

10. Geological Survey of Iran, 1995. Geological Quadrangle Map. NoI11. Geology Organization of Iran.

11. Jafari, A., Finke, P. A., Vande Wauw, J., Ayoubi, S., and Khademi, H. 2012. Spatial prediction of USDA- great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. European Journal of Soil Science, 63(2), 284-298.

12. Jenny Jenny, H. (1941). Factors of Soil Formation: A System of Quantitative Pedology. Mineola.

13. Khaleghi, M., Jafari, A., and Farpour, M. H. 2019. Digital Soil Mapping using legacy soil data: Case study of Faryab region of Kerman. Journal of Agricultural Engineering Soil Science and Agricultural Mechanization,(Scientific Journal of Agriculture), 41(4), 31-48. . (in Persian with English abstract).

14. Landis, J.R., and Koch, G.G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33:159–174.

15. Loeppert, R. H., and Suarez, D. L. 1996. Carbonate and gypsum. Methods of Soil Analysis: Part 3 Chemical Methods, 5, 437-474.

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

17. McBratney, A., Field, D. J., and Koch, A. 2014. The dimensions of soil security. Geoderma,213, 203-213.

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

19. Mousavi S.R., Sarmadian F., Rahmani A., and Khamoushi S.E. 2019. Digital soil mapping with regression classification approaches by RS and Geomorphometrics covariates in the Qazvin plain, Iran. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

20. Nelson, D. W., and Sommers, L. 1983. Total carbon, organic carbon, and organic matter. Methods of Soil Analysis: Part 2 Chemical and Microbiological Properties, 9, 539-579.

21. Omuya, E. O., Okeyo, G. O., and Kimwele, M. W. 2021. Feature selection for classification using principal component analysis and information gain. Expert Systems with Applications, 174, 114765.

22. Pereira, P., Bogunovic, I., Muñoz-Rojas, M., and Brevik, E. C. 2018. Soil ecosystem services, sustainability, valuation and management. Current Opinion in Environmental Science and Health, 5, 7-13.

23. Picard, R. R., and Cook, R. D. 1984. Cross-validation of regression models.Journal of the American Statistical Association, 79(387), 575-583.

24. Rhoades, J. D. 1996. Salinity: Electrical conductivity and total dissolved solids. Methods of Soil Analysis: Part 3 Chemical Methods, 5, 417-435.

25. 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, 110-118.

26. Soil Survey Staff. 2014. Soil Taxonomy: A basic systems of Soil Classification for making and interpreting soil surveys. Twelfth Edition. NRCS. USDA.

27. Sumner, M. E., and Miller, W. P. 1996. Cation exchange capacity and exchange coefficients. Methods of Soil Analysis: Part 3 Chemical Methods, 5, 1201-1229.

28. Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., and Revhaug, I. 2016. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides, 13, 361-378.

29. Wadoux, A. M. C., Minasny, B., and McBratney, A. B. 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.

30. Wei, W., Zhou, B., Połap, D., and Woźniak, M. 2019. A regional adaptive variational PDE model for computed tomography image reconstruction. Pattern Recognition, 92, 64-81.

31. Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., and Finke, P. 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma, 338, 445-452.

32. Zeraatpisheh, M., Garosi, Y., Owliaie, H. R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., and Xu, M. 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. Catena,208, 105723.

33. Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y., Chen, Q., and Shi, Z. 2023. Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. Geoderma, 432, 116383.

34. Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., and Lausch, A. 2021. Prediction of soil organic carbon and the C: N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. Science of the Total Environment, 755, 142661.