

پیش‌بینی مکانی کلاس خاک به کمک مدل درخت تصمیم‌گیری C5.0 تقویت شده با بوستینگ در بخشی از اراضی شهرستان آبیک

- ❖ میلاد ممتازی بروجنی؛ دانش‌آموخته کارشناسی ارشد، گروه علوم و مهندسی خاک، دانشگاه تهران، کرج، ایران
- ❖ فریدون سرمیدیان*؛ استاد گروه علوم و مهندسی خاک، دانشگاه تهران، کرج، ایران.

چکیده

مدیریت منابع خاک بستر اساسی برای حفظ تولیدات جامعه و محیط‌زیست است. خاک‌ها به‌طور معمول برای تولید محصولات کشاورزی و علوفه برای دام‌ها مورد استفاده قرار می‌گیرند. در نتیجه تولید نقشه‌های رقومی با قدرت تفکیک مکانی بالا برای توزیع خصوصیات و کلاس‌های خاک جهت مدیریت خاک و سرزمین بسیار مهم است. مدل درخت تصمیم، یک روش پرکاربرد برای پیش‌بینی کلاس خاک در مطالعات نقشه‌برداری رقومی خاک می‌باشد. هدف از این مطالعه تهیه نقشه‌رقومی خاک در چهار سطح تاکسونومیک با استفاده از مدل درخت تصمیم با الگوریتم C5.0 تقویت‌شده با بوستینگ با کمک داده‌های ماهواره‌ای و مدل رقومی ارتفاع و نقشه‌های زمین‌شناسی و ژئومورفولوژی به عنوان متغیرهای محیطی در ۴۱۰۰۰ هکتار از اراضی شهرستان آبیک می‌باشد. این منطقه با استفاده از روش شبکه‌بندی تصادفی مکان جغرافیایی ۱۲۸ پروفیل خاک مشخص شد و سپس تشریح، نمونه‌برداری و طبقه‌بندی شدند. در این تحقیق با استفاده از روش تحلیل مولفه اصلی بر روی متغیرهای محیطی، در نهایت بیست متغیر محیطی به عنوان نماینده فاکتورهای خاکسازي جهت مدلسازی انتخاب گردید. شاخص همواری دره با درجه تفکیک بالا مهمترین متغیر محیطی می‌باشد که به عنوان ورودی مدل انتخاب گردید. نتایج دقت کلی مدل بوست شده برای پیش‌بینی سطوح تاکسونومیک رده، زیررده، گروه‌بزرگ و زیرگروه به ترتیب ۸۹٪، ۸۵٪، ۵۸٪ و ۵۸٪ نشان داده شد. در این مطالعه همچنین اقدام به بررسی تاثیر روش بوستینگ بر روی مدل درختی گردید، که نتایج نشان داد کلیه سطوح تاکسونومیک با استفاده از مدل بوست شده بهتر از زمانی که از بوستینگ استفاده نشد، پیش‌بینی شدند و بوستینگ سبب بالارفتن دقت کلی و ضریب کاپا گردید.

کلید واژگان: ژئومارفولوژی، ضریب کاپا، متغیر محیطی، مدلسازی، مدل رقومی ارتفاع، نقشه‌برداری رقومی

۱. مقدمه

مدیریت منابع خاک بستر اساسی برای حفظ تولیدات جامعه و محیط زیست است. خاک‌ها به‌طور معمول برای تولید محصولات کشاورزی و علوفه برای دام‌ها مورداستفاده قرار می‌گیرند. در نتیجه تولید نقشه‌های رقومی با قدرت تفکیک مکانی بالا برای توزیع خصوصیات و کلاس‌های خاک جهت مدیریت خاک و سرزمین بسیار مهم است. در همین راستا تحقیقات تهیه نقشه رقومی خاک در سراسر جهان پیشرفت چشمگیری داشته است [۱۷]. که علت آن به‌طور عمده پیشرفت فناوری‌های موجود در روش‌های کمی و سیستم‌های اطلاعات جغرافیایی می‌باشد. در نتیجه روی آوردن به روش‌های جدید در راستای کاربرد بهینه خاک به‌گونه‌ای که هم نیازهای آدمی برآورده شود و هم این منبع تجدید ناپذیر برای آیندگان حفظ شود، ضروری است [۶]. در گذشته نقشه خاک اغلب با روش‌های مرسوم تهیه می‌شد، که از مشکلات روش نقشه‌برداری مرسوم، تفکیک واحدهای خاک بر اساس تفسیر چشمی عکس‌ها، ارتباط دادن لایه‌های مختلف داده‌های محیطی به‌صورت ذهنی، به‌کارگیری علم و دانش تجربی در تجزیه و تحلیل داده‌ها منجر به نادیده گرفته شدن بعضی از واقعیت‌های طبیعی و بروز خطا در به نقشه درآوردن تمام جزئیات می‌شود [۳۲]. با توجه به محدودیت‌ها و کمبودهای موجود در نقشه‌های سنتی، نیاز به استفاده از روش‌های نوین در تهیه نقشه خاک می‌باشد [۳۱]. امروزه مشخص شده که دقت روش‌های پدومتری در برآورد توزیع مکانی داده‌های اندازه‌گیری شده خاک به علت در نظر گرفتن پیوستگی مکانی داده‌ها نسبت به روش‌های معمول آماری بیشتر است [۱۵، ۱۷]. نقشه‌برداری رقومی خاک^۱ روش و مدلی است که بین توزیع خاک (کلاس‌ها یا خصوصیات خاک) و داده‌هایی که به‌آسانی از طریق روش‌های سنجش‌ازدور، تصاویر و عکس‌های ماهواره‌ای و داده‌های ژئومورفومتری به دست می‌آیند و تحت عنوان متغیرهای محیطی نامیده می‌شوند

ارتباط برقرار می‌کند. این داده‌ها به‌عنوان کنترل‌کننده‌ها یا پیش‌بینی‌کننده‌های محیطی که نماینده فاکتورهای خاکسازي هستند نقش دارند [۹]. بنابراین از مدل‌های ریاضی و آماری برای تولید و بهبود نقشه‌های خاک استفاده می‌شود. در همه این مدل‌ها نباید اثر دانش نقشه‌بردار روی کیفیت نتایج نادیده گرفته شود [۱۶]. از دیدگاه مفهومی، روش‌های یادگیری ماشین به‌عنوان بخشی از دانش هوش مصنوعی در عصر فناوری و اطلاعات به‌طور فزاینده‌ای در زمینه‌های مختلف علمی به‌ویژه نقشه‌برداری رقومی خاک توسعه یافته است [۱۷]. طی تحقیقی در استان کرمانشاه به مساحت ۵۰۰۰۰ هکتار، با ۷۸ خاخر اقدام به بررسی توانایی الگوریتم بهینه‌شده درخت تصمیم‌گیری (BDTA) در آشکار ساختن الگوی پراکنش خاک‌ها در سطح گروه بزرگ و زیرگروه خاک شد. جهت مدل‌سازی در این مطالعه به‌منظور پیش‌بینی کلاس‌های خاک از الگوریتم نظارتی درخت تصمیم‌گیری به نام C5.0 که به کمک روش بوستینگ تقویت شده بود استفاده گردید. ارزیابی صحت پیش‌بینی‌ها به‌ترتیب با ضریب کاپا معادل ۰/۷۲ و ۰/۶۶ و صحت عمومی ۰/۸۰ و ۰/۷۶ انجام گرفت [۳۷]. همچنین طی تحقیقی در اراضی استان قزوین به‌وسعت ۶۳۰۰۰ هکتار با استفاده از ۱۵۶ خاخر اقدام به مقایسه مدل‌های مختلف یادگیری ماشینی جهت پیش‌بینی سطح زیرگروه خاک گردیده است، که نتایج آن نشان می‌دهد الگوریتم C5.0 مدل درخت تصمیم با صحت کلی ۵۳٪ و ضریب کاپا ۰/۳۹ نسبت به سایر مدل‌ها دارای صحت بالاتری بوده است [۲۰]. در تحقیق دیگری محققان برزلی اقدام به بررسی الگوریتم درختی C5.0 و رگرسیون لجستیک جهت پیش‌بینی زیررده‌های خاک با ۳۶۱ خاخر در یک منطقه جنگلی گرمسیری برزلی با وسعت ۱۰۲۰۰ هکتار نمودند. نتایج نشان داد که مدل درختی C5.0 با خطای ۴۲ درصد در مرحله ارزیابی نسبت به مدل رگرسیون لجستیک با خطای ۵۲ درصد، به نحو بهتری زیررده‌های خاک را پیش‌بینی می‌کند [۲۵]. هدف از این تحقیق تولید نقشه‌های رقومی

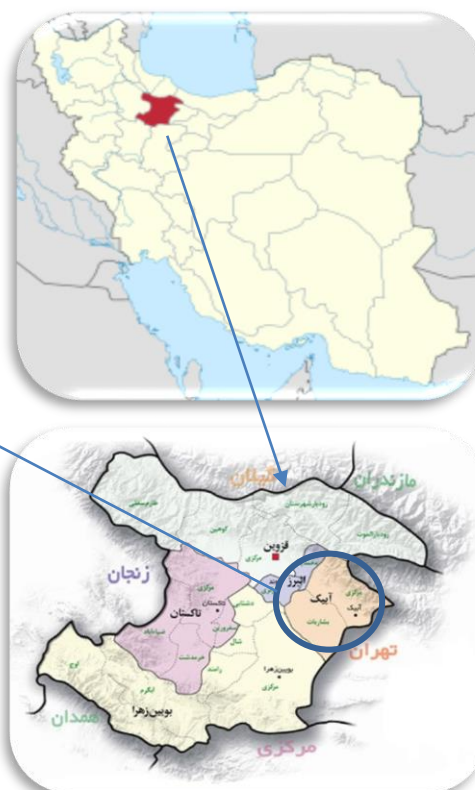
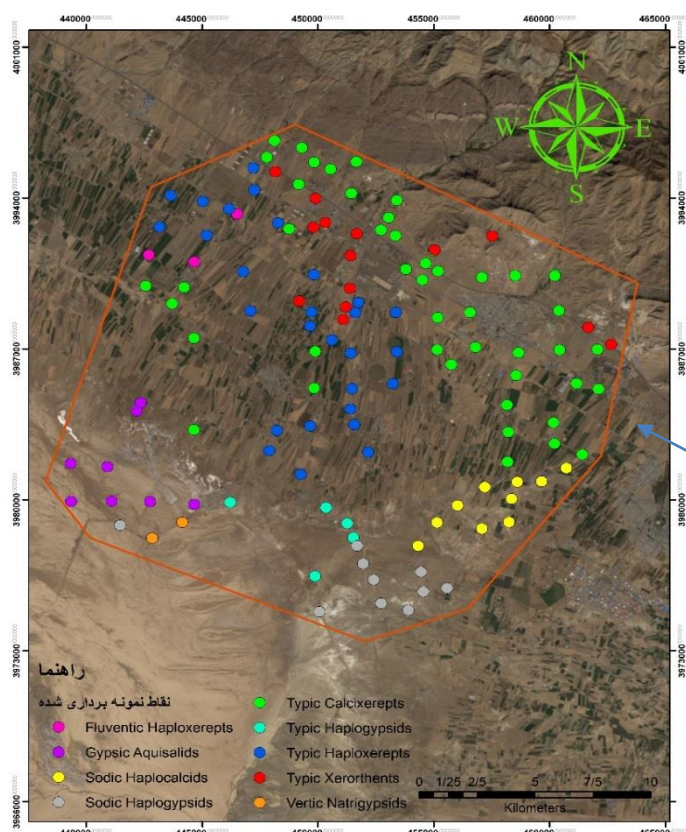
کواترنری می‌باشد. شهرستان آبیک دارای آب‌وهوای معتدل و نیمه خشک است و در سمت جنوب این شهرستان مناطق خشک قرار دارند. به‌منظور تعیین اقلیم منطقه از داده‌های دوره آماری ۲۲ ساله هواشناسی ایستگاه باغ کوثر و ۹ ساله ایستگاه بوئین‌زهر استفاده گردید. که طبق گزارش‌ها، میانگین بارندگی سالیانه در ایستگاه باغ کوثر ۲۷۸ میلی‌متر و ایستگاه بوئین‌زهر ۲۲۲ میلی‌متر و رژیم بارش مدیترانه‌ای است و همچنین متوسط دمای سالیانه در ایستگاه باغ کوثر ۱۳/۸۸ درجه سانتی‌گراد و ایستگاه بوئین‌زهر ۱۶/۸۴ درجه سانتی‌گراد می‌باشد، که در نهایت رژیم حرارتی خاک منطقه موردنظر ترمیک و رژیم رطوبتی آن زیریک خشک و سمت جنوب منطقه مورد مطالعه اریدیک ضعیف و در نواحی پست آکوئیک می‌باشد. گیاهان بومی منطقه مورد مطالعه شامل: گون، سالسولا، گرامینه، کنگر وحشی و غیره می‌باشند. همچنین محصولات زراعی و تحت کشت منطقه بیشتر گندم و ذرت و جو و یونجه هستند.

کلاس‌های رده، زیررده، گروه بزرگ و زیر گروه خاک به کمک نقشه‌ی ژئوform منطقه، با استفاده از مدل درختی، الگوریتم (C5.0) تقویت‌شده با بوس‌تینگ و بررسی تاثیر بوس‌تینگ بر روی صحت مدل در مقایسه با مدل بدون بوس‌تینگ می‌باشد.

۲. مواد و روش‌ها

۱.۲. منطقه مورد مطالعه

منطقه مورد مطالعه به مساحت ۴۱۲۳۹ هکتار در شرق استان قزوین و غرب استان البرز و در محدوده‌ی شهرستان آبیک که از نظر مختصات بین طول جغرافیایی ۴۹°۱۷ تا ۴۵°۴۷'۴۷" و عرض جغرافیایی ۳۹°۴۴'۴۸" تا ۳۹°۸۲'۲۶" متر در زون ۳۹ واقع شده‌است. منطقه به‌صورت کلی دارای زمین‌نماهای کوه و دشت می‌باشد. اغلب تشکیلات زمین شناسی منطقه مربوط به دوره



شکل ۱. منطقه مورد مطالعه و نقاط نمونه‌برداری

۲.۲. مطالعات میدانی و طرح نمونه برداری

براساس مدل اسکورپن [۱۶] دستیابی به نقشه برداری رقومی خاک نیاز به داده های هدف (کلاس خاک) می باشد. لذا در این تحقیق ابتدا توسط ابزار ET GEOWIZARD در نرم افزار آرک جی آی اس، شبکه هایی با فواصل ۱۷۵۰ متر ایجاد شد و محل نقاط نمونه برداری توسط مدل نمونه برداری شبکه ای نامنظم به تعداد ۱۲۸ پروفیل انتخاب شد، و پس از تعیین مکان هندسی نقاط در این نرم افزار، نقاط مورد نظر توسط دستگاه GPS پیمایش شد. سپس در فصول زمستان و بهار حفر پروفیل ها توسط بیل مکانیکی انجام گرفت، همچنین تعداد ۱۲۸ نقطه به صورت مته (به منظور بررسی تغییرات در فاصله میان پروفیل ها) نمونه برداری صورت گرفت. پس از آن، تمام مقاطع خاک بر اساس کتاب (Soil survey staff, 2014) تشریح گردید [۳۰] و از لایه های ژنتیکی کلیه خاک ها، نمونه برداری صورت گرفت و برای تجزیه به آزمایشگاه منتقل گردید. شکل ۱ موقعیت منطقه مورد مطالعه و نقاط نمونه برداری را بر روی تصاویر ماهواره لندست ۸ نشان می دهد.

۳.۲. تجزیه آزمایشگاهی

پس از هوا خشک شدن نمونه ها و عبور از الک دو میلی متری، ویژگی های فیزیکی و شیمیایی نمونه ها شامل بافت خاک، مواد آلی، کربنات کلسیم معادل، گچ، واکنش خاک، هدایت الکتریکی، ظرفیت تبادل کاتیونی، کاتیون ها و آنیون های محلول اندازه گیری شدند. سپس با اتمام آنالیزهای آزمایشگاهی، با استفاده از کتاب Key to Soil Taxonomy با روش طبقه بندی آمریکایی خاک تا سطح فامیلی خاک طبقه بندی صورت گرفت.

۴.۲. متغیرهای محیطی

در تهیه نقشه رقومی خاک بر اساس چارچوب معادله اسکورپن، از اجزای مهم ورودی مدل، متغیرهای

محیطی یا همان نمایندگان فاکتور خاکساز می باشد. در تحقیق حاضر سه نوع متغیر کمکی از جمله مدل ارتفاع رقومی، تصاویر ماهواره ای، نقشه های زمین شناسی و ژئومورفولوژی استفاده گردید. در تحقیق حاضر از مدل رقومی ارتفاع آلوس^۱ در وب سایت جهانی آلاسکا فسیلیتی^۲ تحت نظارت ناسا با قدرت تفکیک مکانی ۱۲/۵ متر استفاده گردیده است. از مدل رقومی ارتفاع جهت استخراج پارامترهای سرزمین از قبیل شیب، جهت شیب، ارتفاع، عمق دره، مساحت حوزه های اصلاح شده، شاخص خیس، شاخص همواری دره با درجه تفکیک بالا (MrVBF) و شاخص بالای پشته با درجه تفکیک بالا (MrRTF) در محیط نرم افزار SAGA GIS استفاده گردید [۹]. همچنین تصاویر ماهواره ای لندست ۸ مربوط به شهرستان آبیگ استان قزوین با قدرت تفکیک مکانی ۳۰ متر جهت پردازش در این تحقیق مورد استفاده قرار گرفت. بر اساس فاکتورهای خاکساز در منطقه، شاخص هایی از قبیل شاخص گیاهی نرمال شده، شاخص شوری، شاخص رس، شاخص کربنات و شاخص روشنایی (نشان دهنده ترکیبات زمین شناسی مواد سطحی) از باندهای تصاویر ماهواره در محیط نرم افزار ERDAS IMAGINE استخراج شد. در نهایت جهت تهیه نقشه ژئوفرم، از نقشه های زمین شناسی سازمان نقشه برداری کشور با مقیاس ۱:۱۰۰۰۰۰ مربوط به مناطق شاکران، قزوین، کرج و اشتهارد جهت استخراج واحدهای لیتولوژی استفاده شد و برای جداسازی لندفرم های منطقه از مدل ارتفاعی رقومی آلوس در نرم افزار ساگا استفاده شد [۹]. همچنین تغییر مقیاس مکانی کلیه شاخص های سنجش از دور و متغیرهای ژئومورفومتری به قدرت تفکیک مکانی یکسان ۳۰ متر در نرم افزار ARC GIS انجام گردید.

۵.۲. تحلیل مؤلفه های اصلی

پس از جمع آوری و بررسی داده های اولیه، تحلیل مؤلفه های اصلی (PCA) در نرم افزار SAS 9.4 انجام

1 ALOS

2 Alaska Facility

در نتیجه یک نقطه پایانی تولید می‌کند. برای محاسبه فاکتور لوودینگ ابتدا لایه‌های مورد استفاده در مطالعه به سه گروه تقسیم شدند. لایه‌های با داده‌ها و محاسبات مشابه به گروه‌هایی از جمله هیدرولوژی سطحی، ارتفاعی و لایه‌های بازتابی تقسیم شدند. گروه هیدرولوژی سطحی شامل شاخص خیزی توپوگرافی^۴، فاکتور LS، جهت جریان، تجمع جریان، نواحی حوزه‌های اصلاح‌شده^۵ و شیب حوزه، طول شیب و عمق جریان می‌باشد. این لایه‌ها می‌تواند به تشخیص عامل محیطی تشکیل خاک همانند مواد مادری و توپوگرافی کمک کند. گروه ارتفاعی شامل جهت، ارتفاع، شیب، عمق دره، شاخص موقعیت توپوگرافی^۶، شاخص همواری دره با درجه تفکیک بالا^۷، شاخص بالای پشته با درجه تفکیک بالا^۸، انحنای نیم‌رخ و انحنای صفحه‌ای است. این لایه عمدتاً ویژگی‌های توپوگرافی از سرزمین را نشان می‌دهد. گروه لایه بازتابی شامل تمام باندهای بازتابی ماهواره موردنظر است که عمدتاً برای نشان پوشش گیاهی و مواد مادری بکار می‌رود [۳۵]. لازم به ذکر است در این تحقیق شاخص‌های ماهواره‌ای را در تحلیل مؤلفه‌های اصلی به علت محدود بودن تعداد شاخص‌ها، وارد نکردیم.

۶.۲. مدلسازی درخت تصمیم

مدل درخت تصمیم، یک روش پرکاربرد برای پیش‌بینی کلاس خاک در مطالعات نقشه‌برداری رقومی خاک می‌باشد. در این مدل‌ها مجموعه داده بر اساس متغیرهای توضیحی برای ایجاد گره‌های پیش‌بینی کننده همگن یا کلاس‌های پیش‌بینی کننده موضوعی [۲۷] و یا خصوصیات خاک‌ها [۲۸] مورد استفاده قرار می‌گیرند. این

گرفت، این کار برای کمک به حذف داده‌های غیرضروری از ۳۱ لایه موجود برای اهداف طبقه‌بندی صورت گرفت. دلیل این کار این است که متغیرهای تکراری بر روی مدل‌ها اثر می‌گذارند زیرا واریانس بسیار ایجاد می‌کنند ولی کارایی ندارند [۲۴]. همچنین گنجاندن تعداد بسیار زیادی متغیرهای توضیحی که دارای همبستگی بالایی هستند، سبب کاهش اعتبار مدل می‌شوند [۲۱]. برای این کار اقدام به محاسبه فاکتور لوودینگ^۱ و بر اساس قدر مطلق آن، حذف متغیر موردنظر صورت گرفت. این فاکتور محاسبه همبستگی بین هر مؤلفه اصلی و متغیر ورودی موردنظر را بر عهده دارد [۸]. در معادله ۱ از مقادیر ویژه و برداری‌های ویژه که حاصل تحلیل مؤلفه اصلی می‌باشد، جهت محاسبه فاکتور لوودینگ استفاده گردید [۱۲].

$$R(kp) = a_{kp} * \sqrt{\lambda p} \quad \text{معادله ۱}$$

$$R(kp) = \text{فاکتور لوودینگ}$$

$$a_{kp} = \text{بردار ویژه}^2 \text{ برای متغیر موردنظر و مؤلفه اصلی موردنظر}$$

$$\sqrt{\lambda p} = \text{مقادیر ویژه}^3 \text{ برای مؤلفه اصلی و واریانس متغیر مورد نظر}$$

همچنین آن تعداد از PCA که ۹۵٪ از واریانس تجعی داده‌ها را شامل می‌شوند، حفظ می‌شوند. قوانین برای این فرایند با حذف داده‌های غیرضروری با عمل تکرار طراحی شده است، که ایجاد یک مجموعه داده جدید در طی هر عمل تکرار همراه با اصلاح می‌کنیم. در نهایت، این کار سبب تولید یک مجموعه داده مرتب می‌شود که تمام مؤلفه‌های اصلی نماینده ۹۵٪ از واریانس هستند و

1 Factor loading

2 Eigen vectors

3 Eigen values

4 Topographic Wetness Index(TWI)

5 Modified Catchment Area(MCA)

6 Topographic Position Index(TPI)

7 Multiresolution index of valley Bottom

Flatness(MrVBF)

8 Multiresolution Ridgetop Flatness index(MrRTF)

تصمیم‌گیری توسعه می‌یابد و هر درخت تلاش می‌کند که نرخ خطای طبقه‌بندی اشتباه را کاهش دهد. سپس هر درخت یک پیش‌بینی ایجاد می‌کند و در نهایت از بین این پیش‌بینی‌ها رأی‌گیری می‌شود. در انتها یک پیش‌بینی با بیشترین تعداد رأی به عنوان پیش‌بینی نهایی انتخاب می‌شود. بوس‌تینگ یک فرا الگوریتم ترکیبی در حوزه یادگیری ماشینی است که برای کاهش عدم توازن و واریانس به کار می‌رود. این روش در یادگیری تحت نظارت مورد استفاده قرار می‌گیرد. این روش برای تبدیل سیستم‌های یادگیری ضعیف به قوی براساس ترکیب نتایج طبقه‌بندی مختلف کاربرد دارد [۱۹]. عملیات مدل‌سازی در نرم‌افزار R 3.3.3 با استفاده از بسته نرم‌افزاری C5.0 صورت گرفت. همچنین برای اعتبار سنجی مدل از بسته‌ی نرم‌افزاری ithir و جهت تولید نقشه به صورت رستری از بسته‌ی نرم‌افزاری Raster استفاده شد [۴]. لازم به ذکر است که ۲۰ درصد از داده‌ها در مرحله آزمون^۱ و ۸۰ درصد داده‌ها در مرحله آموزش^۲ مورد استفاده قرار گرفت.

۷.۲. اعتبار سنجی مدل

۱.۷.۲. ماتریس خطا^۳

نتایج ارزیابی دقت معمولاً به صورت ماتریس خطا ارائه می‌شود که در این صورت انواع پارامترها و مقادیری که بیانگر صحت و یا نوعی خطا در نتایج هستند از این ماتریس استخراج می‌شوند. این ماتریس حاصل مقایسه پیکسل به پیکسل، پیکسل‌های معلوم با پیکسل‌های متناظر در نتایج طبقه‌بندی است. در ماتریس خطا داده‌های زمینی در ستون‌ها و داده‌های مربوط به نتایج طبقه‌بندی در سطرها این ماتریس آورده می‌شود. اعدادی که روی قطر اصلی ماتریس قرار می‌گیرند، تعداد پیکسل‌هایی را مشخص می‌کنند که برچسب آن‌ها در دو سری داده

مدل بر اساس الگوریتم‌های مختلفی مورد استفاده قرار می‌گیرد به نحوی که در سال ۱۹۷۹ محقق الگوریتمی به نام ID3 طراحی کرد که با توجه به ضعف‌هایی که این الگوریتم داشت و سپس در سال ۱۹۹۳ گونه بهبود یافته‌ای از این الگوریتم را با نام C4.5 توسعه داد. این الگوریتم ارباب کمتری داشت و با مقادیر گمشده سازگار بود [۲۳، ۲۲]. پس از گذشت سال‌ها این الگوریتم جای خود را به الگوریتم توسعه یافته‌ای با نام C5.0 داد. C5.0 از نسبت بهره اطلاعات برای ارزیابی تقسیمات در هر گره استفاده می‌کند [۱۳]. در این مطالعه برای استخراج ارتباط خاک-سرزمین و پیش‌بینی الگوی پراکنش خاک‌ها از الگوریتم C5.0 با و بدون روش بوس‌تینگ استفاده شده است. درخت تصمیم‌گیری از بین متغیرهای محیطی، متغیری را برای تشخیص کلاس خاک انتخاب می‌کند که دارای بهره اطلاعاتی (معادله ۳) بالایی باشد. از معیار کاهش آنتروپی (معادله ۲) برای بررسی سودمندی متغیر انتخاب شده استفاده می‌کند [۱].

$$E(t) = \sum p(j|t) \log p(j|t) \quad \text{معادله ۲}$$

$p(j|t)$ فراوانی نسبی کلاس j در گره t می‌باشد. بنابراین بهره اطلاعات برای هر یک از تقسیم‌ها به صورت روبه‌رو به دست می‌آید:

$$GAIN_{split} = E(p) - (\sum_{k=1}^n p(k) E(j|k)) \quad \text{معادله ۳}$$

که در آن گره والد P به اجزاء k تقسیم می‌شود و n_i هم تعداد رکوردهای موجود در جزء i می‌باشد. به این ترتیب شباهت میان داده‌های موجود در یک گره و تفاوت میان مجموعه داده‌های موجود در گره‌ها بیشینه می‌شود. الگوریتم C5.0 از روش هرس بدبینانه برای حذف طبقه‌بندی اشتباه استفاده می‌کند [۲۹]. همچنین با فرا الگوریتم بوس‌تینگ یک توالی از درخت‌های

1 Test data

2 Train data

3 Confusion Matrix

طبقه‌بندی‌شده به مجموع پیکسل‌های یک سطر (کلاس) می‌باشد. این صحت برای کاربر مهم است.

$$U.A = \frac{\sum_{i=1}^n X_{ij}}{\sum_{i=1}^n X_{ij}} \quad \text{معادله ۶}$$

۵.۷.۲. ضریب کاپا^۲

ضریب کاپا دقت طبقه‌بندی را نسبت به یک طبقه‌بندی کاملاً تصادفی محاسبه می‌کند، به این معنی که کاپا دقت طبقه‌بندی را نسبت به حالتی که یک تصویر کاملاً به صورت تصادفی طبقه‌بندی شود نمایش می‌دهد. برای نمونه کاپا معادل ۸۰ درصد یعنی این که نتایج طبقه‌بندی ۸۰ درصد بهتر از موقعی است که پیکسل‌ها به طور تصادفی برچسب دهی شوند. اگر کاپا برابر ۱۰۰ درصد شود به معنی یک طبقه‌بندی کاملاً صحیح بر اساس نمونه‌های گرفته‌شده می‌باشد. مقادیر بین ۰ تا ۱۰۰ درصد هر یک سطح معینی را به نسبت این طبقه‌بندی (کاملاً صحیح) نشان می‌دهند و مقادیر منفی نشان‌دهنده نتایج بسیار بد طبقه‌بندی است. رابطه محاسبه این ضریب به صورت زیر است:

$$k = \frac{N \sum_{i=1}^r x_{ii} - \sum_{i=1}^r (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^r (x_{i+} * x_{+i})} \quad \text{معادله ۷}$$

در این رابطه N تعداد کل پیکسل‌های واقعی زمینی، x_{i+} مجموع عناصر سطر i و x_{+i} مجموع عناصر ستون i می‌باشد. تک پارامترهایی نظیر ضریب کاپا و دقت کلی تنها با کل طبقه‌بندی سروکار دارند و اطلاعات در مورد تک تک کلاس‌ها و یا توزیع مکانی خطاها ارائه نمی‌دهند، برای برآورد دقت کلاس‌ها به صورت مجزا از پارامترهای دیگری نظیر دقت کاربر و دقت تولیدکننده استفاده می‌شود.

هم‌خوانی دارد یا به عبارت دیگر تعداد پیکسل‌هایی که درست طبقه‌بندی شده‌اند روی این قطر قرار می‌گیرند. عناصر غیر قطری مجموعه خطاها می‌باشند. بر مبنای ماتریس خطا پارامترهای متعددی برای بیان صحت و خطا استخراج می‌گردند [۳۴] از آن جمله:

۲.۷.۲. صحت کلی^۱

صحت کلی، میانگینی از دقت طبقه‌بندی است که نسبت پیکسل‌های صحیح طبقه‌بندی‌شده به جمع کل پیکسل‌های معلوم را نشان می‌دهد. که به صورت زیر نمایش داده می‌شود:

$$O.A = \frac{\sum_{i=1}^c E_{ii}}{N} \times 100 \quad \text{معادله ۴}$$

که در این رابطه، C = تعداد کلاس‌ها، N = تعداد کل پیکسل‌های معلوم، E_{ii} = اعضای قطری ماتریس خطا و $O.A$ = صحت کلی طبقه‌بندی یا به نحوی دقیق‌تر میانگین صحت طبقه‌بندی می‌باشد.

۳.۷.۲. صحت تولیدکننده^۲

صحت تولیدکننده احتمال اینکه طبقه‌بندی کننده پیکسلی را به یک کلاس خاص نسبت داده باشد، در صورتی که کلاس واقعی آن مشخص باشد، را بیان می‌کند. به بیان دیگر عنصر قطری هر کلاس به جمع مقادیر هر ستون (هر کلاس) می‌باشد. محاسبه صحت تولیدکننده برای هر کلاس به صورت زیر است:

$$P.A = \frac{X_{ij}}{\sum_{i=1}^n X_{ij}} \quad \text{معادله ۵}$$

۴.۷.۲. صحت کاربر^۳

صحت کاربر احتمال طبقه‌بندی یک کلاس خالص مطابق با همان کلاس در نقشه واقعیت زمینی را بیان می‌کند. به بیان دیگر نسبت پیکسل‌های صحیح

۳. نتایج

۱.۳. شناسایی و رده‌بندی خاک‌ها

پس از طبقه‌بندی مشخص شد که منطقه مورد مطالعه دارای سه نوع خاک در سطح رده، پنج نوع خاک در سطح زیررده، هفت نوع خاک در سطح گروه بزرگ و نه نوع خاک در سطح زیرگروه بزرگ می‌باشد (جدول ۱). با یک دید کلی از نتایج آزمایشگاهی و طبقه‌بندی خاک‌ها می‌توان به این نتیجه رسید که خاک‌های منطقه اغلب دارای میزان بالای آهک (در قسمت میانی و شمالی) و میزان بالای گچ و نمک (در مناطق جنوبی و جنوب شرق) و همچنین ماده آلی پایین می‌باشد، که این‌ها همگی

جدول ۱. طبقه‌بندی خاک‌های منطقه مورد مطالعه تا سطح زیرگروه

زیرگروه (تکرار)	گروه بزرگ (تکرار)	زیررده (تکرار)	رده (تکرار)
Typic Xerorthents (۱۴)	Xerorthents (۱۴)	Orthents (۱۴)	Entisols (۱۴)
Fluventic Haploxerepts (۳)	Haploxerepts (۳۱)	Xerepts (۷۹)	Inceptisols (۷۹)
Typic Haploxerepts (۲۸)	Calcixerepts (۴۸)	Salids (۸)	
Typic Calcixerepts (۴۸)	Aquisalids (۸)	Calcids (۱۰)	
Gypsic Aquisalids (۸)	Haplocalcids (۱۰)		Aridisols (۳۵)
Sodic Haplocalcids (۱۰)	Haplogypsids (۱۵)	Gypsids (۱۷)	
Sodic Haplogypsids (۱۰)	Natrigypsids (۲)		
Typic Haplogypsids (۵)			
Vertic Natrigypsids (۲)			

تغییرات مکانی کلاس‌های خاک منطقه اینگونه است که، خاک‌های تیپیک زراورتنز در مناطق با ارتفاع و شیب بالا (در قسمت شمالی) و یا مناطق تحت فرسایش (شمال غربی) مشاهده می‌شوند. هر دو این مناطق زمان کافی را برای تشکیل افق مشخصه زیرسطحی (انتی سول) به خاک‌ها نمی‌دهد. خاک‌های تیپیک کلسی‌زریپتز و تیپیک هاپلوزریپتز در مناطق میانی منطقه تشکیل شده‌اند، که بیشتر مناطق میانی شمالی و غربی امکان تجمع آهک در افق‌های مشخصه بیشتر وجود داشته و خاک‌های تیپیک کلسی‌زریپتز را به وجود آوردند ولی خاک‌های مناطق میانی و شرقی

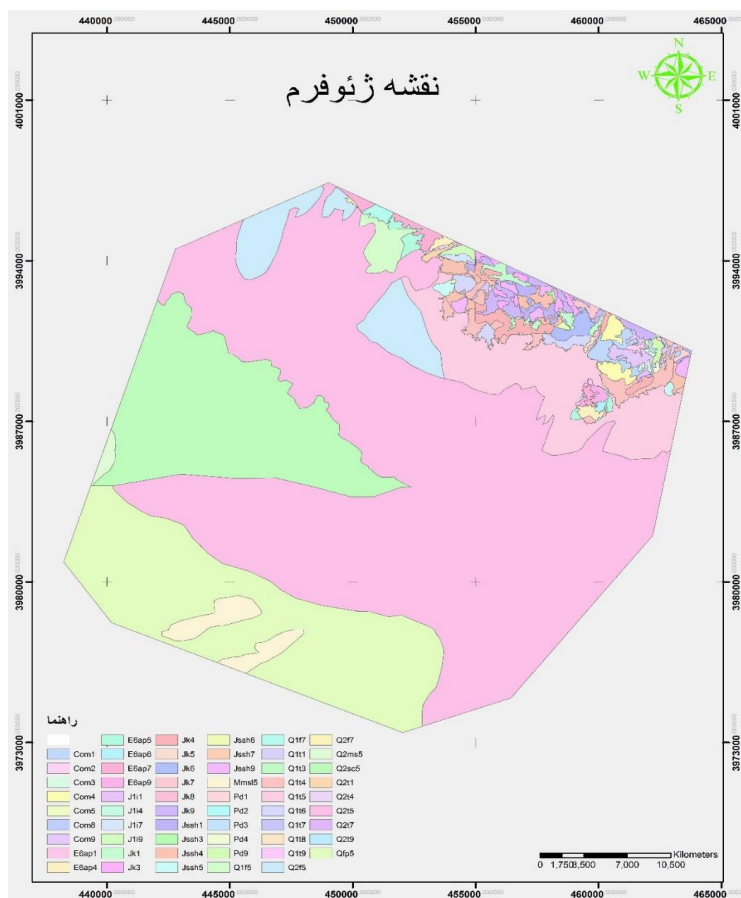
تجمعات آهک به نحوی نیست که افق کلسیک تشکیل گردد و بیشتر افق کمبیک گسترش یافته، خاک‌های تیپیک هاپلوزریپتز تشکیل شده است. در قسمت شمال غربی منطقه، خاک‌های فلوونتیک هاپلوزریپتز وجود دارد که نشان از وجود رسوبات آبرفتی در منطقه دارد. قسمت‌های جنوبی منطقه مورد مطالعه را خاک‌های اریدی سول به علت وجود رژیم رطوبتی اریدیک و افق‌های جیپسیک و سالیک تشکیل داده‌اند. که در جنوب غرب منطقه به علت خاک‌های با میزان EC بالا در همه‌ی افق‌ها، وجود شرایط اکسید و احیا به علت سطح آب زیرزمینی بالا و حضور گچ سبب تشکیل خاک‌های

(لندفرم) حاصل از مدل ارتفاعی‌رقومی که از نرم‌افزار ساگا استخراج شد استفاده گردید. که در نهایت نقشه‌ی ژئوفرم منطقه نمایش دهنده‌ی تغییرات اشکال اراضی و سنگ‌شناسی منطقه می‌باشد. پس از تفکیک واحدها، ۱۴ واحد لیتولوژی در شش دوره زمین‌شناسی در منطقه مورد مطالعه شناسایی شد و نرم‌افزار ساگا توانست منطقه را در ۱۰ لندفرم جداگانه طبقه‌بندی کند. پس از تلفیق این دو نقشه با یکدیگر توانستیم نقشه ژئوفرم منطقه را با ۵۸ واحد ایجاد نماییم (شکل ۲)، که اطلاعات مربوط به هر واحد لیتولوژی و لندفرم را در جداول ۲ و ۳ می‌توان مشاهده نمود. طی مطالعاتی در اراضی شهرستان کوهین از نقشه ژئوفرم جهت تولید نقشه‌ی کلاس‌های خاک استفاده شد [۲]. همچنین محققان دیگری در مطالعات خود از واحدهای ژئومورفیک به عنوان متغیرهای محیطی استفاده نمودند [۳۲، ۱۱].

جیپسیک آکوئی‌سالیید گردید. خاک‌های بعدی ورتیک ناتریجیپسیک می‌باشند که در جنوب منطقه به علت حضور گچ و میزان بالای نسبت سدیم تبادلی و ترک‌ها و شکافت‌هایی که بر روی خاک‌ها ایجاد شده بود، تشکیل شدند. خاک‌های دیگری که در آن مناطق به علت لیتولوژی و وجود مارن‌هایی با لایه‌هایی گچی تشکیل شده‌اند، خاک‌های تیپیک هاپلو جیپسید هستند که تنها دارای افق مشخصه جیپسیک می‌باشند. همچنین در مناطق جنوب غربی به علت خاک‌هایی با افق جیپسیک و کلسیک و نسبت سدیم تبادلی بالا امکان ایجاد سدیک هاپلوکلسیک و سدیک هاپلو جیپسیک را فراهم آورده است.

۲.۳. تهیه نقشه ژئوفرم

برای تهیه نقشه‌ی ژئوفرم از دو نقشه‌ی لیتولوژی حاصل از نقشه‌ی زمین‌شناسی و نقشه‌ی اشکال اراضی



شکل ۲. نقشه ژئوفرم منطقه مورد مطالعه

جدول ۲. اشکال اراضی (لندفرم) منطقه مورد مطالعه

واحد لندفرم	کد	معادل	توصیف واحد لندفرم
Streams	۱	آبراهه	مناطق با خطوط آبراهه‌ای
Midslope drainages	۲	زهکش شیب میانی	زهکشی که در بالای یک دره یا پایین یک قله قرار نگرفته باشد
Upland drainages	۳	زهکش مناطق بالایی	زهکشی که در مناطق بالادستی یک دره قرار دارد
Valleys	۴	دره	فرورفتگی میان دو کوه گفته می‌شود که در یک جهت گسترش یافته است
Plains	۵	دشت	مناطق هموار با شیب بسیار پایین
Open slope	۶	شیب باز	لندفرمی که در پایین یک قله یا منطقه هموار یا بالای دره قرار دارد
Upper slope	۷	شیب بالایی	لندفرمی در پایین یک قله یا منطقه هموار باشد و در بالای دره نباشد
Local ridges	۸	پال‌های محلی	لندفرم هایی در شیب بالایی با قله‌های باریک و طول قله بزرگ‌تر از پهنای لندفرم
Midslope ridges	۹	پال‌های شیب میانی	لندفرم‌هایی در شیب میانی با قله‌های باریک و طول قله بزرگ‌تر از پهنای لندفرم
High ridges	۱۰	پال‌های مرتفع	لندفرم‌ها با ارتفاع زیاد و قله‌های باریک و طول قله بزرگ‌تر از پهنای لندفرم

جدول ۳. واحدهای لیتولوژی منطقه مورد مطالعه

واحد لیتولوژی	دوره زمین‌شناسی	توصیف واحد زمین‌شناسی
Com	(C) کامبرین	سنگ‌آهک، دولومیت، ماسه‌سنگ و شیل،
E6 _{ap}	(E) آئوسین	داسیت، تراکی آندزیت پورفیری خاکستری
J1 _i		سنگ‌آهک کرم روشن به‌طور محلی دولومیت
Jk	(J) ژوراسیک	سنگ‌آهک خاکستری و تیره عمدتاً توده‌ای
JS _{sh}		ماسه‌سنگ آركوزی با بین‌لایه‌هایی از شیل
Mm _{sl}	(M) میوسن	مارن قرمز با میان لایه‌های ژپس و نمک
Pd	(P) پرمین	کوارتزیت و ماسه‌سنگ خاکستری و سنگ‌آهک
Q1 _f		مخروط افکنه های قدیمی
Q1 _t		پادگانه‌های (تراسه) قدیمی
Q2 _f		مخروط افکنه های جوان
Q2 _{ms}	(Q) کواترنری	کفه‌های گلی - نمکی
Q2 _{sc}		کفه‌های سیلتی - رسی
Q2 _t		پادگانه‌های (تراسه) جوان
Qf _p		پهنه‌های سیلابی

۳.۳. تحلیل مؤلفه اصلی

حضور تعداد بسیار زیادی متغیرهای محیطی که دارای همبستگی بالایی هستند، سبب کاهش صحت مدل می‌شود، در نتیجه تحلیل مؤلفه اصلی با روش فاکتور لودینگ انجام گرفت [۲۱]. در اینجا نتایج تحلیل مؤلفه‌های اصلی بر روی باندهای بازتابی در جداول ۴، ۵ و ۶ گزارش شده است. بردارهای ویژه جهت توزیع‌شدگی داده‌ها را برای متغیر و مؤلفه اصلی موردنظر مشخص

می‌کند. همچنین مقادیر ویژه بیانگر رابطه‌ی مؤلفه اصلی و واریانس متغیر مورد نظر می‌باشد. در تمامی جداول، خانه‌های بردارهای ویژه با عددی مشخص شده‌اند که در بازه‌ی منفی ۱ تا مثبت ۱ قرار دارند. هر چه این عدد کمتر باشد به این معنی است که دو پارامتر (در محل تقاطع آن عدد) با یکدیگر ارتباط معکوس دارند و هر چه قدر این عدد بزرگتر باشد یعنی دو پارامتر با یکدیگر همبستگی مثبت دارند. در نهایت فاکتور لودینگ،

در نتیجه این باندها به مرحله دوم از تحلیل مؤلفه اصلی می‌روند. در جدول ۵-الف مشاهده می‌کنید که تا PCA2، مقدار واریانس ۹۵/۳ درصد وجود دارد که شامل ۹۵ درصد از واریانس کل می‌شود. در نتیجه دو لایه باید حفظ شود، آن لایه‌هایی که در جدول ۵-ب دارای بالاترین مجموع فاکتور لوودینگ هستند (باندهای ۱ و ۵) حفظ می‌شوند و به مرحله سوم جهت آنالیز مؤلفه اصلی راه پیدا می‌کنند. که در اینجا تمام مؤلفه‌ها نماینده ۹۵ درصد از واریانس کل شده است و باند ۱ و ۵ برای مدلسازی انتخاب شده است (جدول ۶). به‌طور مشابه محققان آمریکایی [۳۵، ۲۱] از روش آنالیز مؤلفه‌های اصلی با کمک فاکتور لوودینگ جهت انتخاب نهایی پارامترهای باندی بازتابی جهت ورود به مدل استفاده نمودند. شاخص‌های NDVI و DVI جایگزینی برای میزان پوشش گیاهی منطقه می‌باشد [۲۶].

اندازه‌گیری مستقیم از میزان همبستگی یک متغیر با همه مؤلفه‌های اصلی در یک مجموعه داده را فراهم می‌کند. به این ترتیب، متغیرهایی که حداکثر همبستگی را با تمام مؤلفه‌های اصلی دارند، پیشنهاد می‌شوند، زیرا این متغیرها بیشترین اطلاعات یک مجموعه داده معین را نشان می‌دهند. با این تفاسیر، از متغیرهایی که بیشترین مجموع فاکتور لوودینگ در یک مجموعه داده را دارا هستند، برای مدلسازی نهایی استفاده گردید.

پس از آنالیز جدول ۴ مشاهده می‌شود که تا PCA4، مقدار واریانس تجمعی ۹۹ درصد وجود دارد که شامل ۹۵ درصد از واریانس کل می‌شود. در نتیجه با توجه به اطلاعات جدول ۴-الف چهار لایه حفظ گردید، و با توجه به جدول ۴-ب آن لایه‌هایی که دارای بالاترین مجموع فاکتور لوودینگ هستند (باندهای ۷، ۵، ۲، ۱) حفظ می‌شوند.

جدول ۴. تحلیل مؤلفه اصلی باندی-مرحله اول (الف: تحلیل مؤلفه‌ها، ب: فاکتور لوودینگ)

الف-تحلیل مؤلفه اصلی

لایه‌ها	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7
باند ۱	۰/۲۲	-۰/۶۶	۰/۷۰	۰/۰۱۴	۰/۱۱	-۰/۰۲۵	-۰/۰۱
باند ۲	۰/۴۲	۰/۰۲	-۰/۲۲	۰/۴۸	۰/۵۶	-۰/۲۶	۰/۳۹
باند ۳	۰/۴۴	۰/۰۳	-۰/۱۲	۰/۳۴	-۰/۰۷	۰/۰۵۵	-۰/۸۲
باند ۴	۰/۴۴	-۰/۰۷	-۰/۰۸	۰/۲۰	-۰/۷۷	۰/۰۹	۰/۴۰
باند ۵	۰/۲۰	۰/۷۳	۰/۶۲	۰/۰۵	۰/۰۵	۰/۱۶	۰/۰۷
باند ۶	۰/۴۱	۰/۱۱	-۰/۰۳	-۰/۶۲	-۰/۰۴	-۰/۶۵	-۰/۰۷
باند ۷	۰/۴۲	-۰/۰۸	-۰/۲۱	-۰/۴۷	۰/۲۶	۰/۶۸	۰/۰۸
مقادیر ویژه	۴/۹۸	۱/۱	۰/۵۴	۰/۳۲	۰/۰۴۲	۰/۰۲	۰/۰۰۴
واریانس	%۷۱	%۱۵	%۸	%۲	%۲	%۱	%۱
واریانس تجمعی	%۷۱	%۸۶	%۹۴	%۹۶	%۹۸	%۹۹	%۱۰۰

ب-فاکتور لوودینگ

لایه‌ها	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	SUM
باند ۱	۰/۵۰	۰/۷۰	۰/۵۲	۰/۰۱	۰/۰۲	-۰/۰۱	۰/۰۰۱	۱/۷۵
باند ۲	۰/۹۴	۰/۰۲	-۰/۱۶	۰/۲۷	۰/۱۲	-۰/۰۴	۰/۰۲	۱/۵۷
باند ۳	۰/۹۷	۰/۰۳	-۰/۰۹	۰/۱۹	-۰/۰۱	۰/۰۱	-۰/۰۵	۱/۳۶
باند ۴	۰/۹۸	-۰/۰۸	-۰/۰۶	۰/۱۱	-۰/۱۶	۰/۰۱	۰/۰۲۵	۱/۴۲
باند ۵	۰/۴۵	۰/۷۷	۰/۴۶	۰/۰۲۸	۰/۰۱	۰/۰۲	۰/۰۰۴	۱/۷۴
باند ۶	۰/۹۲	۰/۱۱	-۰/۰۲	-۰/۳۵	-۰/۰۱	-۰/۰۹۷	۰/۰۰۴	۱/۵۲
باند ۷	۰/۹۴	-۰/۰۸	-۰/۱۶	-۰/۲۶	۰/۰۵	۰/۱۰۲	۰/۰۰۵	۱/۶۱

به‌طور مشابه، محققانی در تانزانیا [۱۴] و ایران [۳۸، ۱۸] شاخص درخشندگی، شاخص شوری، شاخص رس و شاخص کربنات مورد استفاده قرار دادند. پس از تحلیل مؤلفه اصلی گروه باندهای بازتابی و انتخاب باندهای یک و پنج، این‌بار تمامی این محاسبات برای متغیرهای گروه هیدرولوژی و ارتفاعی صورت گرفت که جداول ۷ و ۸ نتایج نهایی متغیرهای انتخاب شده برای این دو گروه را نشان می‌دهد.

در نتیجه پارامترهای مساحت حوزه‌های اصلاح‌شده، شاخص رطوبتی توپوگرافی، LS فاکتور، تجمع جریان و جهت جریان از گروه هیدرولوژی و پارامترهای شیب،

انحنای صفحه‌ای، انحنای نیم‌رخ، شاخص همواری دره با درجه تفکیک بالا و شاخص بالای پشته با درجه تفکیک بالا از گروه ارتفاعی به مرحله مدلسازی وارد شدند [۳۶، ۳۴، ۱۷]. در رابطه با روش‌های مختلف برای انتخاب متغیرهای نهایی برای ورود به مدل، محققان دیگری از روش انتخاب مبتنی بر همبستگی متغیر [۳۳]، انتخاب بر اساس معیارهای اهمیت متغیر در روش رندوم فارست [۵، ۱۰] و همچنین روشی به نام فیلترکردن ANOVA را به‌عنوان تکنیک‌های کاهش میزان داده‌ها مورد استفاده قرار دادند [۳].

جدول ۵. تحلیل مؤلفه اصلی باندی-مرحله دوم (الف: تحلیل مؤلفه‌ها، ب: فاکتور لوودینگ)

الف- تحلیل مؤلفه اصلی

لایه‌ها	PCA1	PCA2	PCA5	PCA7
باند ۱	۰/۳۸۵۸	-۰/۶۳۵۹	۰/۶۶۶۸	-۰/۰۴۵۶
باند ۲	۰/۶۱۷۲	۰/۰۷۱۶	-۰/۳۳۷	-۰/۷۰۷
باند ۵	۰/۲۹۹	۰/۷۶۷	۰/۵۶۳	۰/۰۷۰۳
باند ۷	۰/۶۱۶۸	-۰/۰۴۶	-۰/۳۵۲۷	۰/۷۰۲
مقادیر ویژه	۲/۲۵۳	۱/۰۷۹	۰/۵۱۲	۰/۱۵۵
واریانس	%۵۳	%۴۲	%۲	%۳
واریانس تجمعی	%۵۳	%۹۵	%۹۷	%۱۰۰

ب- فاکتور لوودینگ

لایه‌ها	PCA1	PCA2	PCA5	PCA7	SUM
باند ۱	۰/۵۷۹۱	-۰/۶۶۰۶	۰/۴۷۷	-۰/۰۱۷۹	۱/۷۴
باند ۲	۰/۹۲۶۶	-۰/۰۷۴۳۶	-۰/۲۴۱۴	-۰/۲۷۸۵	۱/۵۲
باند ۵	۰/۴۴۹۲	۰/۷۹۶۸	۰/۴۰۳۱	۰/۰۲۷۶۹	۱/۶۸
باند ۷	۰/۹۲۵۹	-۰/۰۴۷۸	-۰/۲۵۲۵	۰/۲۷۶۵	۱/۵

جدول ۶. تحلیل مؤلفه اصلی باندی-مرحله سوم (الف: تحلیل مؤلفه‌ها، ب: فاکتور لوودینگ)

الف- تحلیل مؤلفه اصلی

ب- فاکتور لوودینگ

لایه‌ها	PCA1	PCA5	SUM
باند ۱	۰/۷۳	۰/۶۸	۱/۴۱
باند ۵	-۰/۷۳	۰/۶۸	۱/۴۱

لایه‌ها	PCA1	PCA5
باند ۱	۰/۷۰۷	۰/۷۰۷
باند ۵	-۰/۷۰۷	۰/۷۰۷
مقادیر ویژه	۱/۰۷۴	۰/۹۲۵
واریانس	%۷۴	%۲۶
واریانس تجمعی	%۷۴	%۱۰۰

جدول ۷. تحلیل مؤلفه اصلی گروه هیدرولوژی

لایه‌ها	PCA1	PCA2	PCA4	PCA4	PCA5
MCA	۰/۳۹	۰/۵۰	-۰/۴۷	-۰/۳۵	۰/۴۹
TWI	-۰/۴۰	۰/۵۵	-۰/۰۵	۰/۷۰	۰/۲۱
LS-F	۰/۵۶	-۰/۱۱	۰/۵۹	۰/۳۱	۰/۴۷
Flow-Acc	۰/۵۰	۰/۴۹	۰/۱۱	۰/۱۲	-۰/۷۰
Flow-dir	-۰/۳۶	۰/۴۴	۰/۶۳	-۰/۵۲	۰/۰۷
مقادیر ویژه	۱/۵۸	۱/۴	۰/۸۷	۰/۶۱	۰/۵۲
واریانس	%۳۲	%۲۷	%۱۸	%۱۲	%۱۱
واریانس تجمعی	%۳۲	%۵۹	%۷۷	%۸۹	%۱۰۰

جدول ۸. تحلیل مؤلفه اصلی گروه ارتفاعی

لایه‌ها	PCA1	PCA2	PCA3	PCA4	PCA5
Mrrtf	۰/۵۴	-۰/۰۰۰۶	۰/۷۷	-۰/۱۲	-۰/۳۲
Mrvbf	۰/۵۷	-۰/۲۹	-۰/۰۶	۰/۱۵	۰/۷۵
انحنای صفحه‌ای	۰/۲۰	۰/۶۷	-۰/۰۵	۰/۷۱	-۰/۰۴
انحنای نیم‌رخ	۰/۲۴	۰/۶۴	-۰/۲۰	-۰/۶۷	۰/۱۸
شیب	-۰/۵۳	۰/۲۴	۰/۶۰	۰/۰۰۳	۰/۵۴
مقادیر ویژه	۱/۷۵	۱/۴۵	۰/۷۲	۰/۵۵	۰/۵۲
واریانس	%۳۵	%۲۹	%۱۴	%۱۱	%۱۱
واریانس تجمعی	%۳۵	%۶۴	%۷۸	%۸۹	%۱۰۰

۴.۳. تهیه نقشه رقومی خاک

در نهایت اقدام به پیش‌بینی کلاس‌های خاک در سطوح رده، زیررده، گروه بزرگ و زیرگروه با استفاده از الگوریتم C5.0 درخت تصمیم که به‌وسیله‌ی روش بوستینگ تقویت شده است، گردید (شکل ۳، ۴، ۵، ۶). برای تقویت مدل در مرحله‌ی یادگیری و آموزش (۸۰٪ از داده‌ها) از روش بوستینگ بهره بردیم که سبب از بین رفتن خطاهای طبقه‌بندی‌های غلط در مرحله آموزش شد و با ساختن تعداد درخت‌هایی بین ۱۰۰ تا عدد، دقت کلی و ضریب کاپا را به ترتیب به ۱۰۰٪ و ۱ رساند. که در این حالت آموزشی همه‌ی کلاس‌های خاک در سطوح مختلف رده‌بندی به دقتی معادل ۱۰۰ درصد رسیدند. سپس این مدل درختی را بر روی داده‌های آزمون (۲۰٪ از داده‌ها)، آزمایش کرده تا دقت این مدل برای پیش‌بینی کلاس‌های سطح رده خاک مشخص شود. همچنین نتایج

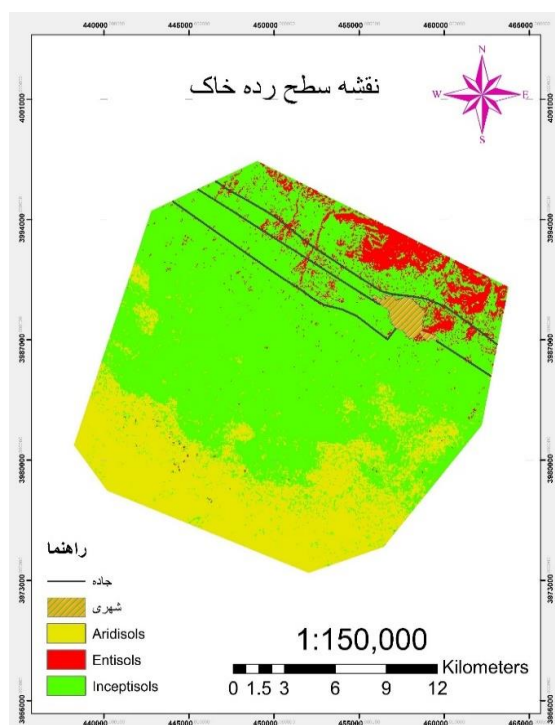
مربوط به ماتریس خطا برای مرحله آزمون در جداول ۹،

۱۰، ۱۱، ۱۲ ارائه شده است.

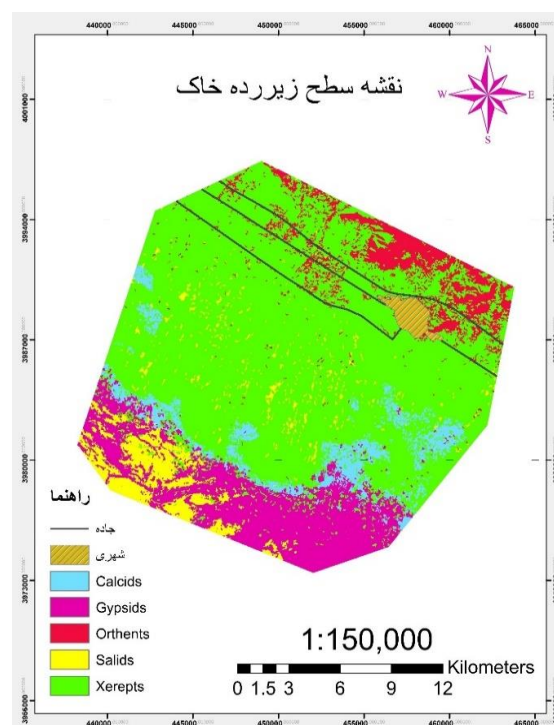
در این گروه داده، بالاترین دقت در سطح رده مربوط به کلاس انتی سول با دقت تولیدکننده ۱۰۰ درصد و ضعیف‌ترین مربوط به اربدی سول با دقت ۷۵ درصد می‌باشد. در سطح زیررده بالاترین دقت مربوط به کلاس سالید با دقت تولیدکننده ۱۰۰ درصد و ضعیف‌ترین دقت مربوط به کلاس اورتنز با دقت ۵۰ درصد می‌باشد. همچنین در سطح گروه بزرگ بالاترین دقت‌ها مربوط به کلاس ناتری جیپسید و هاپلوجیپسید با دقت تولیدکننده ۱۰۰ درصد و ۸۰ درصد و ضعیف‌ترین مربوط به کلاس زراورتنز با دقت ۲۵ درصد می‌باشد. در نهایت در سطح زیرگروه، بالاترین دقت مربوط به کلاس سدیک هاپلوکلسید و سدیک هاپلوجیپسید با دقت تولیدکننده ۱۰۰ درصد و ضعیف‌ترین تیپیک هاپلوجیپسید با دقت ۳۴ درصد می‌باشد. برای

باند ۱ بازتابی و جهت جریان مهمترین متغیرهای محیطی هستند که توسط مدل استفاده شدند.

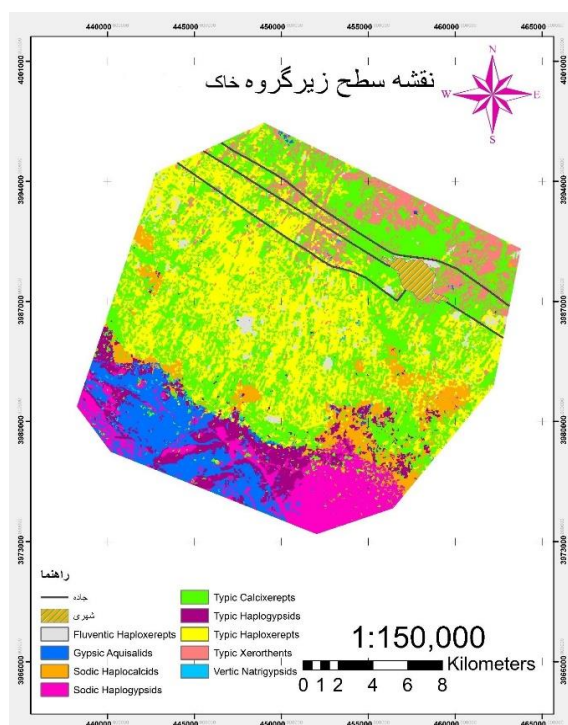
ساخت درخت جهت پیش بینی کلاس های خاک، شاخص همواری دره با درجه تفکیک بالا، شاخص پوشش گیاهی،



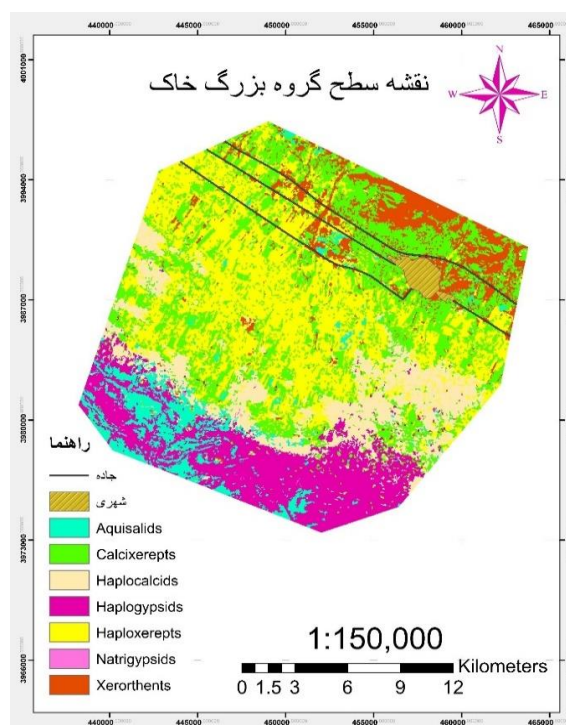
شکل ۴. نقشه رقومی سطح زیربرده



شکل ۳. نقشه رقومی سطح رده خاک



شکل ۶. نقشه رقومی سطح زیرگروه خاک



شکل ۵. نقشه رقومی سطح گروه بزرگ خاک

جدول ۹. نتایج ماتریس خطا مربوط به الگوریتم C5.0 بوست شده در مرحله آزمون (سطح رده خاک)

دقت کاربر	اینسیتی سول	انتی سول	اریدی سول	
٪۱۰۰	۰	۰	۶	اریدی سول
٪۵۰	۱	۱	۰	انتی سول
٪۸۹	۱۶	۰	۲	اینسیتی سول
	٪۹۵	٪۱۰۰	٪۷۵	دقت تولید کننده

جدول ۱۰. نتایج ماتریس خطا مربوط به الگوریتم C5.0 بوست شده در مرحله آزمون (سطح زیررده خاک)

دقت کاربر	زریتر	سالید	اورتنز	جیپسید	کلسید	
٪۰	۱	۰	۰	۰	۰	کلسید
٪۸۰	۰	۰		۴	۱	جیپسید
٪۱۰۰	۰	۰	۱	۰	۰	اورتنز
٪۱۰۰	۰	۱	۰	۰	۰	سالید
٪۸۸	۲۱	۰	۱	۱	۱	زریتر
	٪۹۵	٪۱۰۰	٪۵۰	٪۸۰	٪۰	دقت تولید کننده

جدول ۱۱. نتایج ماتریس خطا مربوط به الگوریتم C5.0 بوست شده در مرحله آزمون (سطح گروه بزرگ خاک)

دقت کاربر	چ	ج	ث	ت	پ	ب	آ	
٪۱۰۰	۰	۰	۰	۰	۰	۰	۲	آ-آکوئی سالید
٪۵۰	۱	۰	۲	۱	۰	۴	۰	ب-کلسی زریتر
٪۰	۰	۰	۱	۰	۰	۱	۰	پ-هاپلوکلسید
٪۸۰	۰	۰	۰	۴	۰	۰	۱	ت-هاپلو جیپسید
٪۵۰	۲	۰	۳	۰	۰	۱	۰	ث-هاپلو زریتر
٪۱۰۰	۰	۱	۰	۰	۰	۰	۰	ج-ناتری جیپسید
٪۵۰	۱	۰	۰	۰	۰	۱	۰	چ-زراورتنز
	٪۲۵	٪۱۰۰	٪۵۰	٪۸۰	٪۰	٪۵۸	٪۶۷	دقت تولید کننده

جدول ۱۲. نتایج ماتریس خطا مربوط به الگوریتم C5.0 بوست شده در مرحله آزمون (سطح زیرگروه خاک)

دقت کاربر	خ	ح	چ	ج	ث	ت	پ	ب	آ	
٪۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	آ-فلوونتیک هاپلو زریتر
٪۱۰۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	ب-جیپسید آکوئی سالید
٪۱۰۰	۰	۰	۰	۰	۰	۰	۱	۰	۰	پ-سدیک هاپلوکلسید
٪۵۰	۱	۰	۰	۰	۰	۱	۰	۰	۰	ت-سدیک هاپلو جیپسید
٪۷۰	۰	۰	۱	۰	۷	۰	۰	۲	۰	ث-تیپیک کلسی زریتر
٪۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	ج-تیپیک هاپلو جیپسید
٪۵۸	۰	۱	۴	۰	۲	۰	۰	۰	۰	چ-تیپیک هاپلو زریتر
٪۲۰	۰	۱	۱	۰	۳	۰	۰	۰	۰	ح-تیپیک زراورتنز
٪۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	خ-ورتیک ناتری جیپسید
	٪۰	٪۵۰	٪۶۷	٪۰	٪۵۹	٪۱۰۰	٪۱۰۰	٪۳۴	٪۰	دقت تولید کننده

در نهایت اقدام به مقایسه‌ی مدل‌های ایجادشده برای سطوح مختلف رده‌بندی به‌وسیله‌ی الگوریتم C5.0 به‌صورت تقویت‌شده با روش بوستینگ و تقویت نشده با بوستینگ گردید. اساس این مقایسه استفاده از دقت کلی و ضریب کاپا برای هر سطح رده‌بندی در هر دو روش قبل و بعد از بوستینگ می‌باشد. به‌نحوی که از نتایج جدول ۱۳

قابل فهم می‌باشد، در کلیه سطوح رده‌بندی دقت کلی و ضریب کاپای حاصل از اعتبار سنجی مدل بعد از تقویت به‌وسیله‌ی بوستینگ بالاتر رفته است. در نتیجه استفاده از روش بوستینگ بر روی مدل، عاملی برای بهبود دقت نقشه‌های ایجادشده می‌باشد.

جدول ۱۳. مقایسه مدل‌های تقویت شده و نشده توسط بوستینگ در سطوح مختلف رده‌بندی

سطح تاکسونومیک	شاخص اعتبارسنجی	بدون بوستینگ	با بوستینگ
رده	دقت کلی	۸۱٪	۸۹٪
رده	ضریب کاپا	۰/۶۴	۰/۷۶
زیر رده	دقت کلی	۷۷٪	۸۵٪
زیر رده	ضریب کاپا	۰/۴۷	۰/۶۶
گروه بزرگ	دقت کلی	۵۴٪	۵۸٪
گروه بزرگ	ضریب کاپا	۰/۳۷	۰/۴۷
زیر گروه	دقت کلی	۵۰٪	۵۸٪
زیر گروه	ضریب کاپا	۰/۳۵	۰/۴۳

۴. بحث و نتیجه‌گیری

با استفاده از الگوریتم C5.0 مدل درخت تصمیم اقدام به پیش‌بینی خاک‌های کل منطقه گردید، که نتایج حاکی از دقت کلی ۸۹ درصد برای سطح رده، ۸۵ درصد برای سطح زیررده، ۵۸ درصد برای سطح گروه بزرگ و زیرگروه بزرگ بوده است. به‌نحوی که می‌توان بیان نمود الگوریتم C5.0 کلیه سطوح را با دقت مناسبی پیش‌بینی کرده است. نتایج تحقیق مشابهی در استان قزوین با استفاده از مدل تقویت شده درخت تصمیم نشان از دقت کلی ۴۹٪ و ضریب کاپا ۰/۳۳ برای پیش‌بینی سطح زیرگروه خاک دارد [۲۰]. همچنین نتایج پیش‌بینی کلاس‌های سطح گروه بزرگ و زیر گروه خاک توسط مدل درختی C5.0 و تقویت شده با بوستینگ در استان کرمانشاه نشان از ضریب کاپای به ترتیب ۰/۷۲ و ۰/۶۶ دارد [۳۷]. همچنین پژوهشگران دانمارکی جهت پیش‌بینی کلاس‌های خاک از مدل درختی بوست شده استفاده نمودند که نتایج پژوهش آنها دقت کلی ۵۱٪ را نشان داد [۱۹]. همچنین در این

تحقیق از آنالیز مؤلفه‌های اصلی با احتساب فاکتور لوودینگ برای حذف داده‌های غیرضروری و تکراری استفاده شد که توانست ۱۱ متغیر محیطی که دارای داده‌های تکراری و متشابه با دیگر متغیرها بودند را حذف نماید، به‌نحوی که این متغیرهای تکراری اگر وارد مدل شوند سبب بالارفتن خطا بر روی عملکرد مدل می‌شود. این نوع روش تحلیل مؤلفه اصلی در تحقیق مشابهی در آریزونا آمریکا انجام گرفته است که متغیرهای انتخابی سبب ایجاد پیش‌بینی کلاس خاک با دقت بالا شده است [۳۵]. در فرآیند نقشه‌برداری رقومی کلاس‌های خاک، مدل درخت تصمیم C5.0 به‌خوبی توانسته ارتباط بین داده‌های خاک و متغیرهای محیطی را مدل کند. اجرای روش بوستینگ بر روی مدل به‌خوبی توانست مدل را تقویت نماید و سبب بالارفتن دقت پیش‌بینی‌های مدل گردد. که در این راستا در سال‌های گذشته محققینی از مدل درخت تصمیم با الگوریتم C5.0 که به‌وسیله بوستینگ تقویت شده است، جهت نقشه‌برداری رقومی خاک استفاده نمودند که نتایج قابل قبول و مشابهی ارائه

تحقیق می‌توان به این اشاره نمود که تکنیک درخت تصمیم در نقشه‌برداری رقومی خاک می‌تواند نقشه‌برداری خاک‌ها را در گستره‌ای وسیع و متشکل از هرگونه عوارض طبیعی به یکباره انجام داده و سبب بهبود نقشه‌های تولیدشده نسب به نقشه‌های سنتی گردیده است و سبب بالا رفتن سرعت انتقال داده‌ها و اطلاعات در نقشه‌ها شده است. به‌خوبی نقشه‌برداری رقومی توانست بر انتقادات وارد بر نقشه‌برداری سنتی فائق آید، و به‌خوبی توانست سبب ایجاد نقشه‌های پیوسته از داده‌ها و اطلاعات خاک در تمام منطقه مورد مطالعه گردد. به‌طور کلی محققین مختلفی کارایی یادگیری ماشینی و مدل درخت تصمیم را در پیش‌بینی کلاس‌های خاک تأیید کرده‌اند و در مطالعات خود از این مدل استفاده نموده‌اند [۳۸، ۳۲، ۳۱]. از آنجاییکه الگوریتم C5.0 بوسه‌شده بر روی مدل درخت تصمیم به‌خوبی توانست ارتباط بین متغیرهای محیطی و کلاس‌های خاک را ایجاد نماید و از طرفی کاربرد این الگوریتم در زمینه تهیه نقشه خاک محدود می‌باشد، در نتیجه پیشنهاد می‌گردد که محققان در مطالعات آینده جهت نقشه‌برداری رقومی خاک اقدام به استفاده از این الگوریتم و مقایسه‌ی آن با دیگر الگوریتم‌های یادگیری ماشینی نمایند.

گردیده است [۳۷، ۳۶، ۲۵، ۱۹]. از طرفی یک معیار مهم در دقت کلی مدل درخت تصمیم برای پیش‌بینی کلاس‌های خاک، تعداد داده‌های خاک مشاهداتی (پروفیل‌های خاک) می‌باشد، به‌نحوی که در سطح زیرگروه کلاس تیپیک هاپلوزپتز که دارای مشاهدات بالایی بودند با دقت بالاتری پیش‌بینی شدند. در حقیقت تنوع و غیریکنواختی خاک‌ها یکی از عواملی است که باعث پیچیدگی و دشواری نقشه‌برداری رقومی می‌شود، به‌نحوی که هرچه تنوع و تغییرات کلاس‌های خاک بیشتر باشد، تشخیص و تمایز انواع کلاس‌های خاک توسط مدل‌ها به سختی انجام می‌گیرد. در نتیجه اجرای تکنیک نقشه‌برداری رقومی و برای رسیدن به نتایج دقیق، نیاز به اطلاعات ورودی دقیق (اطلاعات پروفیلی و متغیرهای محیطی مناسب) با تراکم مناسب از منطقه مورد مطالعه دارد. همچنین در مورد متغیرهای محیطی مؤثر حاصل از مدل ارتفاعی رقومی می‌توان از شاخص همواری دره با درجه تفکیک بالا و جهت جریان، و متغیرهای محیطی حاصل از تصاویر ماهواره باندهای بازتابی ۱ و ۵ و شاخص پوشش گیاهی را به‌عنوان مؤثرترین پارامترها در پیش‌بینی کلاس‌های خاک دانست. که استفاده از این پارامترها را می‌توان در مطالعات مشابهی که صورت گرفته است، مشاهده نمود [۳۳، ۱۸، ۱۷، ۱۱]. از نتایج دیگر این

References

- [1] Adu-Poku, S. (2012). Comparing classification algorithms in data mining. MSc dissertation. Central Connecticut State University.
- [2] Alijani, Z., Sarmadian, F. and Mousavi, R. (2014). Comparing the Accuracy of Soil Map Prepared by Geopedology and Usual Method of Iran (Case Study: Kouhin). Journal of range and watershed management, 67, 93-102.
- [3] Behrens, T., Zhu, A., Schmidt, K. and Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, 155, 175-185.
- [4] Beucher, A., Møller, A.B., and Greve, M.H. (2017). Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. Geoderma, 352, 351-359.
- [5] Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A. and Edwards, Jr. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 240, 68-83.

- [6] Daigle, J.J., Hudnall, W.H., Gabriel, W.J., Mersiovsky, E. and Nielson, R.D. (2005). The National Soil Information System (NASIS): Designing soil interpretation classes for military land-use predictions. *Journal of terra mechanics*, 42(3), 305-320.
- [7] Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, 152(3), 195-207.
- [8] Harris, R.J. (2001). *A Primer of Multivariate Statistics*. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.
- [9] Hengl, T., Rossiter, D.G. and Stein, A. (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Soil Research*, 41(8), 1403-1422.
- [10] Heung, B., Bulmer, C.E. and Schmidt, M.G. (2014). Predictive soil parent material mapping at a regional-scale: a random forest approach. *Geoderma*, 214-215, 141-154.
- [11] Jafari, A., Finke, P.A., Van deWauw, 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. *Eur. J. Soil Sci*, 63, 284-298.
- [12] Jensen, J.R. (2005). *Introductory Digital Image Processing a Remote Sensing Perspective*. Prentice Hall Series in Geographic Information Science. Pearson Prentice Hall, Upper Saddle River, New Jersey, 526 pp.
- [13] Kuhn, M., Weston, S., Coulter, N., Culp, M. and Quinlan, J.R. (2015). Package 'C50'. <https://cran.rproject.org/web/packages/C50/C50.pdf> (accessed 20-07-16).
- [14] Massawe, B.H., Subburayalu, S.K., Kaaya, A.K., Winowiecki, L. and Slater, B.K. (2016). Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. *Geoderma*, 311, 143-148.
- [15] McBratney, A.B., Odeh, I.O., Bishop, T.F., Dunbar, M.S. and Shatar, T.M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma*, 97(3), 293-327.
- [16] McBratney, A.B., Mendonc. M.L., Santos. A. and Minasny. B. (2003). On digital soil mapping. *Geoderma*, 117, 3-52.
- [17] Minasny, B. and McBratney, A.B. (2016). Digital soil mapping: a brief history and some lessons. *Geoderma*, 264, 301-311.
- [18] Mirakzehi, K., Pahlavan-Rad, M.R., Shahriari, A. and Bameri, A. (2018). Digital soil mapping of deltaic soils: A case of study from Hirmand (Helmand) river delta. *Geoderma*, 313, 233-240.
- [19] Møller, A.B., Iversen, B.V., Beucher, A. and Greve, M.H. (2017). Prediction of soil drainage classes in Denmark by means of decision tree classification. *Geoderma*, 352, 314-322.
- [20] 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, 415-422.
- [21] Pires, J.C.M., Martins, F.G., Sousa, S.I.V., Alvim-Ferraz, M.C.M. and Pereira, M.C. (2008). Selection and validation of parameters in multiple linear and principal component regressions. *Environmental Modelling & Software*, 23(1), 50-55.
- [22] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- [23] Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the micro electronic age*. Edinburgh University Press.
- [24] Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth A Course in Methods of Data Analysis*. Duxbury Thomson Learning, Pacific Grove, CA, 742 pp.
- [25] Ricardo, O. D., Gustavo, M.V., Maurício, R.C. and Nelson, F.F. (2105). Digital soil mapping for soil class prediction in a dry forest of Minas Gerais, Brazil. *Xv congresso brasileiro de ciencia do solo*, 1-4.
- [26] Rouse Jr, J., Haas, R.H., Schell, J.A. and Deering, D.W. (1974). Monitoring vegetation systems in the Great Plains with ERTS.
- [27] Schmidt, K., Behrens, T. and Scholten, T. (2008). Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. *Geoderma*, 146(1), 138-146.

- [28] Scull, P., Franklin, J. and Chadwick, O.A. (2005). The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling*, 181(1), 1-15.
- [29] Shirali, R. (2016). Classification Trees and Rule-Based Modeling Using the C5. 0 Algorithm for Self-Image Across Sex and Race in St. Louis.
- [30] Soil Survey Staff. (2014). Keys to Soil Taxonomy (12th. ed). USDA-Natural Resources Conservation Service. National Soil Survey Center, Lincoln, NE.
- [31] Subburayalu, S.K., Jenhani, I. and Slater, B.K. (2014). Disaggregation of component soil series on an Ohio county soil survey map using possibilistic decision trees. *Geoderma*, 213, 334-345.
- [32] Taghizadeh-Mehrjardi, R., Minasny, B., McBratney, A.B., Triantafilis, J., Sarmadian, F. and Toomanian, N. (2012). Digital soil mapping of soil classes using decision trees in central Iran. In *Proceedings of the 5th Global Workshop on Digital Soil Mapping*, pp. 197-202.
- [33] Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B. and Triantafilis, J. (2015). Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma*, 253-254, 67-77.
- [34] Taghizadeh-Mehrjardi, R., Sarmadian, F., Minasny, B., Triantafilis, J. and Omid, M. (2014). Digital mapping of soil classes using decision tree and auxiliary data in the Ardakan region, Iran. *Arid Land Research and Management*, 28(2), 147-168.
- [35] Travis, Nauman. (2009). Digital soil-landscape classification for soil survey using Aster satellite and digital elevation data in organ pipe cactus national monument. Msc dissertation in University of Arizona. 1-170.
- [36] Wu, W., Li, A.D., He, X.H., Ma, R., Liu, H.B. and Lv, J.K. (2018). A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China. *Computers and Electronics in Agriculture*, 144, 86-93.
- [37] Zandi, B.M. and Shekaari, P. (2019). Soil Distribution Pattern Analysis in a Low Relief Area Using Decision Trees Algorithm. *Iranian Journal of Soil and Water Research*, 50(2), 463-480.
- [38] Zeraatpisheh, M., Ayoubi, S., Jafari, A. and Finke, P. (2017). Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Journal of Geomorphology*, 285, 186-204.
- [39] Zhu, A.X., Hudson, B., Burt, J., Lubich, K. and Simonson, D. (2001). Soil mapping using GIS, expert knowledge, and fuzzy logic. *Journal of Soil Science Society of America*, 65(5), 1463-1472.

Spatial prediction of soil classes using C5.0 boosted decision tree model in part of Abyek lands

- ❖ **Milad Momtazi Burojeni**; M. Sc. Graduate, Department of Soil Science, University of Tehran, Karaj, Iran.
- ❖ **Fereydoon Sarmadian***; Professor, Department of Soil Science, University of Tehran, Karaj, Iran.

Abstract

Soil resource management is essential to maintain community production and environment. Soil is usually used to produce agricultural products and livestock fodder. As a result, the production of high-resolution spatial digital maps is crucial for the distribution of soil and soil properties and land management. The decision tree model is a widely used method for predicting soil class in digital soil mapping studies. The purpose of this study was to provide a digital soil mapping in four levels of taxonomy using decision tree with Boost-reinforced C5.0 algorithm using satellite data and digital Elevation Model and geological maps as environmental variables in 41,000 hectares of Abyek city. This area was identified using randomized gridding of geographic location of 128 soil profiles and then described, sampled and classified. In this research, using the principal component analysis method on environmental variables, 20 environmental variables were selected as the representative of stacking factors for modeling. Multi resolution Valley Flatness Index is the most important environmental variable that was selected as input of the model. The results of the overall accuracy of the integrated model for predicting taxonomic levels of the Order, Suborder, great group, Subgroup were shown to be 89%, 85%, 58%, 58%, respectively. The study also examined the effect of the boosting technique on the tree model, which showed that all taxonomic levels were better predicted by using the boost model than when no boosting was used, and boosting resulted in an increase in overall accuracy and kappa coefficient it turned out.

Keywords: Digital Elevation Model, Digital mapping, environmental variable, Geomorphology, kappa coefficient, Modeling