مجله علمی‌پژوهشی مهندسی اکوسیستم بیابان
سال هفتم، شماره نوزدهم، تابستان ۱۳۹۷، صفحه ۸۱ـ۹۵

کاربرد الگوریتم‌های داده‌کاوی در تفکیک منابع رسوبی حوزهٔ آبخیز نوده گناباد

مهدی بشیری[1*]، مهسا آریاپور[2]، علی گل‌کاریان[3]

تاریخ دریافت: ۱۳۹۶/۱۲/۱٤ **تاریخ پذیرش: ۱۳۹۷/٤/۱۱**

چکیده

لازمهٔ اجرای برنامه‌های کنترل رسوب، شناسایی اهمیت نسبی منابع رسوب، میزان مشارکت آن‌ها و در نتیجه شناسایی مناطق بحرانی آبخیزهاست. در این پژوهش از الگوریتم‌های داده‌کاوی برای تفکیک منابع رسوبی حوضهٔ نوده گناباد در استان خراسان رضوی با کمک متغیرهای ژئوشیمیایی، دانه‌بندی و سنگ‌شناسی استفاده شد. یازده الگوریتم برای طبقه‌بندی در نرم‌افزار MATLAB برنامه‌نویسی و نتایج براساس ضریب تبیین و میانگین مربع خطا با یکدیگر مقایسه شد. بررسی غلظت عناصر ژئوشیمیایی در هفت واحد زمین‌شناسی حوضه نشان داد که عناصر Ca، Fe، Mg و AL دارای بیشترین و عناصر B و Co دارای کمترین غلظت در نمونه‌های خاک است. ارزیابی کلی الگوریتم‌های طبقه‌بندی در مرحلهٔ آموزش نشان داد که الگوریتم‌های تحلیل ممیزی، جنگل تصادفی، k نزدیک‌ترین همسایه و ماشین‌های بردارپشتیبان با توابع خطی، چندجمله‌ای، چندگانه و شعاع مبنا با حداکثر مقدار ضریب تبیین (R^2=۱) و حداقل مقدار میانگین مربع خطا (MSE=۰)، دقیق‌ترین الگوریتم‌ها در تفکیک منابع رسوبی هستند و روش درخت رگرسیونی ضعیف‌ترین عملکرد را دارد. در مرحلهٔ آزمون نیز ماشین‌های بردارپشتیبان با تابع شعاع مبنا، دقیق‌ترین الگوریتم و درخت طبقه‌بندی با بالاترین خطا، ناکارآمدترین الگوریتم بود. همچنین ورود متغیرهای ژئوشیمیایی منجر به بالاترین دقت در تفکیک منابع رسوبی شد و متغیرهای دانه‌بندی کمترین دقت تفکیک را باعث شد.

واژه‌های کلیدی: الگوریتم‌های طبقه‌بندی، منشأیابی، حوضهٔ نوده، غلظت عناصر.

۱. استادیار گروه مرتع و آبخیزداری، دانشکده کشاورزی و منابع طبیعی، دانشگاه تربت حیدریه، خراسان رضوی/ m.bashiri@torbath.ac.ir
۲. کارشناس ارشد آبخیزداری، دانشکده کشاورزی و منابع طبیعی، دانشگاه تربت حیدریه، خراسان رضوی
۳. استادیار گروه مرتع و آبخیزداری، دانشکده منابع طبیعی و محیط زیست، دانشگاه فردوسی مشهد، خراسان رضوی

DOI: 10.22052/deej.2018.7.19.49

مقدمه

فرسایش خاک پس از رشد جمعیت، دومین چالش مهم زیست‌محیطی جهان است که به‌دلیل داشتن اثرات چندجانبهٔ آشکار و پنهان زیست‌محیطی و اجتماعی، یکی از فرایندهای پیچیده و خطرساز محیطی است. فرسایش خاک و حمل رسوب پدیده‌ای است که در نواحی مختلف و در مقیاس‌های متفاوتی اتفاق می‌افتد (عرب‌خدری، ۲۰۱٤). شناخت عوامل مؤثر در افزایش تخریب خاک به‌منظور درک صحیح‌تر فرایند فرسایش و تخریب خاک در تمامی اقالیم، به‌منظور جلوگیری و کاهش روند تخریب امری ضروری است (تورنبل[1] و همکاران، ۲۰۰۸). در این میان شناخت منشأ رسوب از جمله سهم زیرحوضه‌ها، گامی اساسی برای سیاست‌گذاری مدیریت حوزه‌های آبخیز است. تعیین این سهم کاربرد مؤثری در مدل‌سازی، شناخت الگوهای پراکنش، تدوین روش‌های مهار رسوب و تخمین میزان فرسایش خاک دارد. بدین منظور یافتن روش‌هایی برای برآورد علمی و دقیق‌تر میزان فرسایش و رسوب در حوزه‌های آبخیز کشور امری ضروری است. از نظر ماهیت حرکت، رسوبات به دو شکل بار معلق و بستر انتقال می‌یابند و از نظر منبع پیدایش نیز در دو گروه بار رسوبی حوضه‌ای و آبراهه‌ای تقسیم می‌شوند. آگاهی از مقدار و تغییرات مقادیر رسوب در تنظیم تغییرات بستر، کیفیت آب، رفتار هیدرولیکی و انجام پروژه‌های حفاظت آب و خاک و مدیریت و برنامه‌ریزی آن کمک شایانی می‌کند (میلهوس[2]، ۱۹۸۸). امروزه حجم زیاد داده‌های ذخیره‌شده و نیز گستردگی ابعاد آن سبب شده است که در بسیاری از موارد، روش‌های آماری به‌تنهایی قادر به کشف خصوصیات داده‌ها نباشند. داده‌کاوی[3] به‌عنوان یک راه‌حل برای این مسائل بوده و عبارت است از استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه دادهٔ بسیار بزرگ که این الگوها و دانش‌ها معمولاً مستتر در داده‌ها هستند (چان و لویز[4]، ۲۰۰۲). داده‌کاوی در اواخر دههٔ ۱۹۸۰ پدیدار شد و سپس در دههٔ ۱۹۹۰ گام‌های بلندی در این شاخه از علم برداشته شد و انتظار می‌رود همچنان به رشد و پیشرفت خود ادامه دهد. تاکنون مطالعات مختلفی در زمینهٔ مدل‌سازی فرسایش خاک و تولید رواناب و رسوب با استفاده از الگوریتم‌های داده‌کاوی صورت گرفته است. با توجه به اینکه هدف از مطالعهٔ حاضر، تفکیک و طبقه‌بندی منابع رسوبی حوضه و ارزیابی عملکرد الگوریتم‌های طبقه‌بندی در شناسایی منابع رسوبی با استفاده از ردیاب‌های طبیعی حوضه است، در ادامه به تحقیقات مرتبط با پژوهش حاضر اشاره می‌شود.

حرما[5] و همکاران (۲۰۱۵) مدل‌سازی رواناب و رسوب را در حوضه‌ای واقع در نپال به‌وسیلهٔ شبکه‌های عصبی مصنوعی انجام دادند. نتایج ایشان نشان داد شبکهٔ عصبی مصنوعی[6] کارایی بسیار خوبی در این زمینه دارد. کومارگویال[7] (۲۰۱٤) طی پژوهشی عملکرد مدل درخت M5 و روش رگرسیونی[8] را در مقایسه با شبکهٔ عصبی مصنوعی برای پیش‌بینی آورد رسوبی حوزهٔ آبخیز ناگوا در هند مورد بررسی قرار داد. نتایج نشان‌دهندهٔ برتری عملکرد مدل M5 و رگرسیونی نسبت به روش دیگر بود. حدادچی[9] و همکاران (۲۰۱۳) در مناطق آبرفتی، به اجرای تکنیک انگشت‌نگاری رسوب با استفاده از ردیاب‌ها پرداختند. در این مطالعه، منابع رسوبی با استفاده از هفت مدل آمیخته، ارزیابی و با کمک الگوریتم ژنتیک بهینه‌سازی گردید. نتایج نشان داد که دامنهٔ خطا در تمامی مدل‌ها در محدودهٔ قابل قبول قرار دارد. دمیرسی و بالتاسی[10] (۲۰۱۳) طی پژوهشی در ساکری منتوفی پورت[11] آمریکا، از روش‌های منطق فازی، رگرسیون و منحنی سنجهٔ رسوب برای برآورد مقدار غلظت رسوب معلق و دمای آب که به‌طور پیوسته در مدت زمان پنج سال تهیه شده بود، استفاده کردند. نتایج حاکی از برتری دقت روش منطق فازی نسبت به سایر روش‌ها بود. میسرا[12] و همکاران (۲۰۰۹)، با هدف شبیه‌سازی رواناب و رسوب با استفاده از دو روش ماشین بردارپشتیبان و

1. Turnbull
2. Milhouse
3. Data mining
4. Chan and Lewis

5. Harma
6. Artificial Neural Network
7. Kumar goyal
8. Regression Method
9. Haddadchi
10. Demirci and Baltaci
11. Sacremento Feeport
12. Misra

شبکهٔ عصبی به مطالعهٔ حوضه‌ای در هند پرداختند. در این مطالعه، سری‌های زمانی داده به مجموعه‌های آموزش، واسنجی و اعتبارسنجی تقسیم شد و عملکرد مدل‌ها با استفاده از ضریب همبستگی بررسی گردید. نتایج نشان داد که ماشین بردارپشتیبان بهبود قابل توجهی در آموزش، کالیبراسیون و اعتبارسنجی در مقایسه با مدل شبکهٔ عصبی دارد و می‌تواند به‌عنوان جایگزین مناسبی در برآورد رواناب و پیش‌بینی رسوب مورد استفاده قرار گیرد.

حیات‌زاده و همکاران (۲۰۱۵) طی پژوهشی به پیش‌بینی بار رسوبی حوضهٔ باغ‌عباس با استفاده از دو روش شبکهٔ عصبی و روش‌های رگرسیونی پرداختند. در این تحقیق، داده‌های ۱۳۶ واقعهٔ دبی جریان و پارامترهای مورفولوژیکی مورد تحلیل قرار گرفت. نتایج نشان داد بهترین روش برای برآورد رسوب حوضه، روش شبکهٔ عصبی به همراه داده‌های ژئومورفولوژیکی حوضه است. کاکائی لفدانی (۲۰۱۳) در پژوهشی، توانایی مدل‌های شبکهٔ عصبی و ماشین بردارپشتیبان[1] در برآورد رسوب معلق روزانهٔ رودخانهٔ دویرج واقع در غرب ایران، بررسی و با روش‌های رگرسیونی مقایسه کردند. نتایج حاکی از برتری مدل‌های یادشده نسبت به مدل‌های رگرسیونی بود. جودی و ستاری (۱۳۹۵) در مقایسه برآورد رسوبات رودخانه‌ای نشان دادند روش‌های داده‌کاوی عملکرد بهتری را نسبت به روش‌های سنتی و رگرسیونی نشان دادند. در پژوهش آن‌ها روش‌های سنتی تفاوت چشمگیری نداشته و در برآورد بار رسوبی کل عملکرد مشابهی را داشتند.

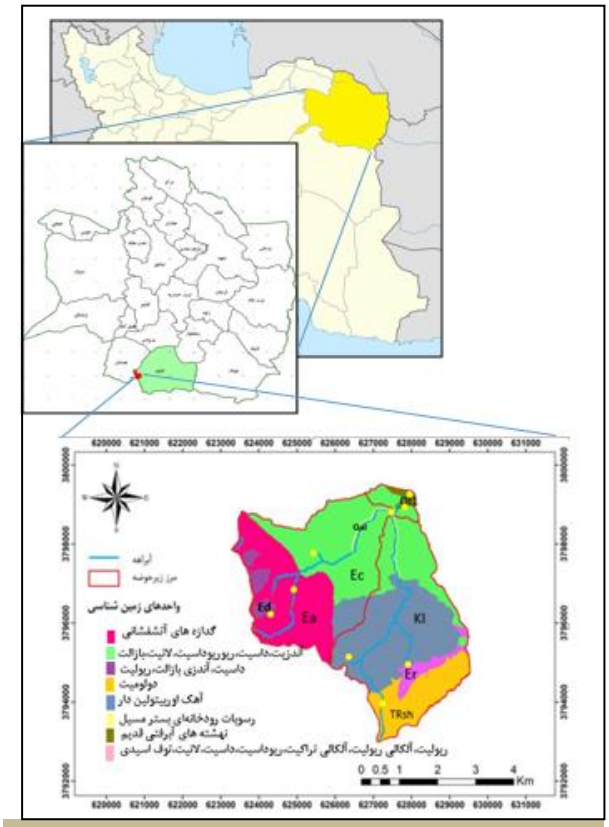
بسالت‌پور و همکاران (۲۰۱۶) در بررسی اثرات ویژگی‌های خاک بر فرسایش‌پذیری حوضهٔ کارون به این نتیجه رسیدند عملکرد مدل SVM نسبت به مدل رگرسیون خطی که حتی قادر به تشخیص روابط و اثرات فاکتورهای موثر بر فرسایش حوضه نبوده، بسیار مناسب معرفی شده است. با توجه به نتایج پژوهش‌های انجام‌شده با روش‌های مختلف در زمینهٔ فرسایش و رسوبات حوضه می‌توان نتیجه گرفت الگوریتم‌های داده‌کاوی در اکثر موارد، دقت بالاتر و عملکرد بهتری را نسبت به سایر روش‌های آماری و سنتی ارائه کرده است. گرچه با توجه به مرور منابع، عمدتاً روش‌های داده کاوی در زمینهٔ مدل‌سازی رسوب به کار رفته‌اند، تکنیک‌های طبقه‌بندی رسوبات با استفاده از الگوریتم‌های موجود، چندان توسعه پیدا نکرده‌اند و در زمینهٔ طبقه‌بندی منابع رسوبی به کاربرد روش‌های سنتی مانند رگرسیون و نهایتاً آنالیز تشخیص اکتفا شده است. لذا این پژوهش با هدف تفکیک منابع رسوبی تجمع پیداکرده در پشت سد خاکی حوزهٔ آبخیز نوده گناباد (خروجی حوضه) با استفاده از علم نوین داده‌کاوی صورت گرفت تا بتواند در زمینهٔ منشأیابی رسوب و تعیین بهترین الگوریتم در این زمینه مورد کاربرد قرار گیرد؛ چراکه این الگوریتم‌ها در زمینه‌های مختلف عملکرد موفقی داشته‌اند. بنابراین انتظار می‌رود که در تفکیک منابع رسوبی نیز موفق عمل نمایند.

مواد و روش‌ها

حوزهٔ آبخیز نوده با مساحتی در حدود ۲۲۹۸/۳۹ هکتار در حوزهٔ آبخیز کویر مرکزی و از نظر کشوری در محدودهٔ شهرستان گناباد در استان خراسان رضوی واقع شده است. موقعیت جغرافیایی حوزهٔ آبخیز نوده بین طول جغرافیایی ″۲۴ ′۲۰ °۵۸ تا ″۳۸ ′۲۴ °۵۸ و عرض جغرافیایی ″۱۳ ′۱۶ °۳۴ تا ″۵۰ ′۲۰ °۳۴ است. این حوضه دارای هفت واحد زمین‌شناسی است که شامل نهشته‌های آبرفتی قدیمی (Qtl)، ریولیت، آلکالی ریولیت، آلکالی تراکیت، ریوداسیت، داسیت، لاتیت، توف اسید (Er)، دولومیت (Tr.sh)، آهک اوربیتولین‌دار(Kl)، داسیت، آندزی بازالت، ریولیت (Ed)، آندزیت، داسیت، ریوداسیت، لاتیت، بازالت (Ec) و گدازه‌های آتشفشانی (Ea) است. همچنین میزان بارندگی سالانهٔ آن ۲۵۴ میلی‌متر است و در خروجی این حوضه، یک سد خاکی احداث شده است. روش کار این پژوهش شامل دو بخش آنالیزهای آزمایشگاهی و آماری است. از آنجایی که داده‌های ژئوشیمیایی رسوبات برای مطالعهٔ آن‌ها از دید رسوبی و زیست‌محیطی بسیار حائز اهمیت است و ترکیب شیمیایی این رسوبات حاکی از پایایی عوامل مختلف زمین‌شناسی مانند جایگاه زمین‌ساختی، ترکیب سنگ منشأ، شدت هوازدگی، بلوغ بافتی و کانی‌شناسی طی حمل و رسوب‌گذاری است

1. Support Vector Machine

(سایگل[1]، ۲۰۰۲)، در بخش آنالیزهای آزمایشگاهی پژوهش حاضر به شناسایی و تعیین غلظت عناصر موجود در نمونه‌های خاک پرداخته شد. شکل (۱) موقعیت جغرافیایی حوزهٔ آبخیز مورد مطالعه به همراه مکان‌های نمونه‌برداری در واحدهای زمین‌شناسی را نشان می‌دهد.



شکل (۱): موقعیت جغرافیایی حوضه و مکان‌های نمونه‌برداری

Figure (1): Geographic location of the watershed and sampling points

در این پژوهش، نمونه‌برداری از واحدهای زمین‌شناسی درون آبراهه‌هایی که تنها از همان واحد سرچشمه می‌گرفتند، انجام شد. بعد از مشخص کردن واحدهای زمین‌شناسی، برای هر تکرار بین ۲۵۰ تا ۵۰۰ گرم از رسوبات رودخانه‌ای تهیه شد. نمونه‌برداری از عمق ۰ تا ۵ سانتی‌متری (فیض‌نیا، ۲۰۰۸؛ روتون و همکاران، ۲۰۱۱) نقاط منتخب و همچنین رسوبات خروجی حوزه به‌صورت سیستماتیک تصادفی و در هر نقطه ۵ تکرار برای هر نمونه انجام شد. در مجموع از ۸ نقطهٔ نمونه‌برداری (۷ واحد و خروجی) ۴۰ نمونه برداشت شد که ۳۵ نمونه مربوط به واحدهای زمین‌شناسی (۵ تکرار برای ۷ واحد) و ۵ نمونه مربوط به خروجی حوضه است.

نمونه‌برداری از واحدهای زمین‌شناسی درون آبراهه‌هایی که فقط از همان واحد سرچشمه می‌گرفتند، انجام شد. در نهایت از هر مجموعهٔ ۵ تایی نمونه واحدهای زمین‌شناسی، ٤ تکرار به‌صورت دوبه‌دو پس از حمل به آزمایشگاه با یکدیگر مخلوط شدند که از ٤٠ نمونهٔ برداشت‌شده، ۲۲ نمونه (یک نمونهٔ خروجی و ۲۱ نمونه منابع رسوب) باقی ماند. برای بررسی دانه‌بندی و بافت خاک از هر نمونه، دو زیرنمونهٔ ۲۰۰ گرمی و ۵۰ گرمی تهیه شد. ابتدا زیرنمونه‌های ۲۰۰ گرمی از هفت الک با منافذ درشت‌تر از ٦٣ میکرون سری الک ادن ــ ونت‌ورث[2] به روش الک‌تر عبور داده شد. همچنین زیرنمونه‌های ۵۰ گرمی که به روش الک خشک از رسوبات زیر ۲ میلی‌متر تهیه شده بود، با استفاده از هیدرومتر و روش ASTM[3] (۲۰۰۸) برای اندازه‌گیری درصد رس، سیلت و شن استفاده شد و سپس پارامترهای دانه‌بندی ذرات رسوب، با استفاده از نرم‌افزار GRADISTAT نسخهٔ ۸ تعیین گردید. برای مطالعات مورفوسکوپی و سنگ‌شناسی در نمونه‌های رسوب، ابتدا سنگ‌های شاخص واحدهای مختلف به روش هدف‌دار (جمع‌آوری خرده‌سنگ‌های شاخص واحدهای کاری) از داخل نمونه‌ها جمع‌آوری شد و بعد از دسته‌بندی (از لحاظ نوع سنگ و کانی‌های موجود در آن‌ها) و شناسایی نوع سنگ، از هریک از نمونه رسوبات واحدها و خروجی حوزه یک زیرنمونه شامل ۱۰۰ عدد خرده‌سنگ به روش تصادفی روی الک با قطر منفذ ٦٣ میکرون تهیه و اقدام به تطابق آن‌ها با سنگ‌های شناسایی شده در زیر بینوکولر[4] شد.

در مرحلهٔ بعد، برای بررسی غلظت عناصر موجود در نمونه‌ها ابتدا از هر نمونه، یک زیرنمونهٔ یک گرمی از رسوبات زیر الک ٦٣ میکرون به روش الک خشک تهیه شد و بعد از هضم اسیدی، به حجم رسانیدن و عبور از کاغذ صافی (یاپ[5] و همکاران، ۲۰۰۲)، با دستگاه ICP OES[6] غلظت عناصر اندازه‌گیری شد. برای این منظور ابتدا ۳۲ عنصر در نمونه رسوبات مخزن سد اندازه‌گیری و سپس عناصری که در

1. Siegel

2. Udden-Wentworth Sieve Series
3. American Society for Testing and Materials
4. Binocular
5. Yap
6. Inductively Coupled Plasma Optical Emission Spectrometry

رسوبات مخزن فاقد مقدار بود حذف شد و در نهایت برای تمامی نمونه‌ها ۲۱ عنصر شامل B، Cr، Co، Cr، Cu، K، Li، Na، P، Pb، Ti، Ca، Al، Mg، Sr، Mn، Ba، Zn، Y، V و Fe اندازه‌گیری گردید. بعد از جمع‌آوری و ثبت داده‌ها، تحلیل‌های آماری پژوهش با استفاده از نرم‌افزارهای MATLAB و EXCEL۲۰۱۳ صورت گرفت. آنالیزهای آماری تحقیق شامل ورود داده‌های حاصل به محیط نرم‌افزار MATLAB و سپس اجرا و ارزیابی الگوریتم‌های به کار رفته است. داده‌های ورودی به الگوریتم‌ها شامل پارامترهای دانه‌بندی و بافت‌سنجی رسوبات، غلظت عناصر شیمیایی و نهایتاً فراوانی سنگ‌ها می‌باشند که همگی به‌صورت کمی وارد آنالیز گردیدند. شایان ذکر است که علاوه بر نقطهٔ خروجی، در ۲۱ نقطه از سطح حوضه، تعداد ۲۱ عنصر شیمیایی، ۱۶ پارامتر بافت‌سنجی و دانه‌بندی و نیز درصد فراوانی ۱۱ واحد سنگ‌شناسی شاخص برآورد شدند. لذا مجموعاً ۱۰۵۶ داده (۴۸ متغیر در ۲۲ نقطه) در آنالیز مورد استفاده قرار گرفت. الگوریتم‌های مورد استفاده در این تحقیق شامل جنگل تصادفی[1]، درخت‌های طبقه‌بندی[2]، شبکهٔ عصبی مصنوعی، تحلیل‌ممیزی[3]، رگرسیون لجستیک[4]، انواع ماشین بردارپشتیبان، روش‌های رگرسیونی، روش گروهی مدل‌سازی داده‌ها[5]، درخت خطی محلی[6]، روش تشخیص الگو[7] و k نزدیک‌ترین همسایه[8] بودند که برای معرفی هریک از این الگوریتم‌ها در ادامه مختصر توضیحاتی آورده شده است.

۱. روش درخت تصمیم: درخت‌ها در هوش مصنوعی[9] برای نمایش مفاهیم مختلفی نظیر ساختار، جملات، معادلات و... استفاده می‌شوند. درخت‌های تصمیم روشی برای نمایش یک سری از قوانین هستند که منتهی به یک رده می‌شوند. این درختان نمونه‌ها را به‌نحوی دسته‌بندی می‌کنند که از ریشه به‌سمت پایین رشد می‌کند و در نهایت به گره‌های برگ می‌رسد. درخت تصمیم شامل تعدادی از الگوریتم‌ها مثل ID 3،C4،C5 و CHAID در طبقه‌بندی است (بریمن[10]، ۲۰۰۱).

۲. روش جنگل تصادفی[11]: این روش یک تکنیک مدرن ناپارامتری[12] و متعلق به خانوادهٔ روش‌های دسته‌جمعی است. در حال حاضر یکی از بهترین الگوریتم‌های یادگیری است و برای بسیاری از مجموعه داده‌ها، دسته‌بندی را با صحت بالایی انجام می‌دهد و برخلاف مدل‌های کلاسیک چون رگرسیون[13] که تنها بر پایهٔ یک مدل تکیه دارند با استفاده از صدها و هزاران درخت از اطلاعات بیشتری در داده‌ها استفاده می‌کند تا بتوان استنباط بهتری از متغیرها داشت. این الگوریتم از جمله دسته‌بندهایی است که متد Bagging را به کار می‌گیرد و حاوی چندین درخت تصمیم است که خروجی آن از خروجی‌های درخت‌های انفرادی به دست می‌آید (همان).

۳. روش‌های رگرسیونی[14]: مدل‌های رگرسیون بر اساس این نظریه ساخته شده‌اند که اگر دو عامل به یکدیگر بستگی داشته باشند، تغییر یکی با تغییر دیگری قرین خواهد شد. هر قدر ارتباط دو عامل به یکدیگر نزدیک‌تر و قوی‌تر باشد، ضریب همبستگی تغییرات آن‌ها بزرگ‌تر و به یک نزدیک‌تر خواهد شد. این روش انواع مختلفی دارد که مهم‌ترین آن‌ها روش رگرسیون خطی[15]، روش رگرسیون درجه‌دوم خالص[16]، و روش رگرسیون اثر متقابل[17] است (بریمن و همکاران، ۱۹۸٤).

٤. روش تحلیل ممیزی (آنالیز تشخیصی)[18]: تحلیل ممیزی تکنیکی چندمتغیره است که با جدا کردن مجموعه‌های متمایز مشاهده‌ها و با تخصیص مشاهده‌های جدید به دسته‌های از پیش تعریف‌شده سروکار دارد. آنالیز تشخیصی برای طبقه‌بندی پاسخگویان بر اساس کدهای یک متغیر وابستهٔ اسمی دو یا چندوجهی به کار می‌رود. تحلیل تشخیصی ترکیب دو یا چند متغیر مستقل را که به بهترین وجه تفاوت بین دو

1. Random Forest
2. Classification Trees
3. Discriminant Analysis
4. Logistic Regression
5. Group Method of Data Handling
6 .Local Linear Model Tree
7. Pattern Recognition Method
8. K Nearest Neighbor
9. Artificial Intelligence
10. Breiman
11. Random Forest
12. Nonlinear
13. Regression
14. Regression Method
15. linear Regression
16. Purequadratic Regression
17. Interaction Regression
18. Discriminant analysis

گروه را تبیین می‌کند، نشان می‌دهد. هدف کلی تحلیل ممیزی به وجود آوردن یک ترکیب خطی بین متغیرهاست که از آن برای گروه‌بندی افراد استفاده شود. این ترکیب خطی، یک مسئلهٔ پیچیده و چندمتغیره را به یک مسئلهٔ آماری ساده و یک متغیره تبدیل می‌کند (پینتو[1] و همکاران، ۲۰۱۲).

۵. مدل درخت خطی محلی[2]: مدل درخت خطی محلی که در آن از نوعی مدل فازی عصبی خطی محلی استفاده شده است، الگوریتمی بر اساس استراتژی تقسیم و حل می‌باشد که در آن حل مسئله پیچیده از طریق تقسیم به تعدادی زیرمسئله کوچک‌تر صورت می‌پذیرد. بنابراین مشخصات این مدل‌های فازی عصبی به مقدار زیادی، به ساختار الگوریتم به‌کاربرده‌شده جهت تقسیم‌بندی وابسته است. این الگوریتم برای رسیدن به خروجی بهتر (خروجی با خطای کمتر) فضای مسئله را به تعدادی مدل خطی محلی یا LLM تقسیم می‌کند و پس از پیدا کردن بدترین LLM (LLM با خطای بیشتر) با تقسیم آن به دو LLM، الگوریتم را ادامه می‌دهد تا به انتها برسد (کشاورز امامی و کارولوکس، ۲۰۰۷).

٦. روش ماشین‌های بردارپشتیبان[3]: ماشین بردارپشتیبان یکی از روش‌های یادگیری با نظارت[4] است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. این روش نسبتاً جدید است و در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر طبقه‌بندی نشان داده است. در این روش با استفاده از همهٔ باندها و یک الگوریتم بهینه‌سازی، نمونه‌هایی که مرز کلاس‌ها را تشکیل می‌دهند، به دست می‌آید و با استفاده از آن‌ها یک مرز تصمیم‌گیری خطی بهینه[5] برای جدا کردن کلاس‌ها محاسبه می‌شود؛ این نمونه‌ها را بردارهای پشتیبان می‌گویند. این روش چند هستهٔ مختلف را به‌طور پیش‌فرض پشتیبانی می‌کند که شامل هسته‌های خطی[6]، چندجمله‌ای[7]، شعاع مبنا[8]، تانژانت هیپربولیک[9]، لاپلاسی[10]، بسل[11]، تابع شعاع مبنای آنوا[12] و اسپلاین[13] هستند. از مزایای ماشین بردارپشتیبان، توانایی حل مسائل طبقه‌بندی پیچیده با تعداد لایه‌های زیاد و نمونه‌های آموزشی کم است (اورال بنا[14] وهمکاران، ۲۰۱۰).

۷. روش آنالیز k نزدیک‌ترین همسایه (KNN)[15]: آنالیز نزدیک‌ترین همسایه[16] یکی از روش‌های طبقه‌بندی متغیرها بر اساس تشابه آن‌ها با یکدیگر است که الگوی داده‌ها را بدون نیاز به الگوهای از پیش مشخص، طبقه‌بندی می‌نماید. در این روش سعی می‌شود تا ویژگی‌های نقاط داده از روی ویژگی‌های نزدیک‌ترین همسایگانشان تعیین شوند. یکی از بهترین طبقه‌بندی‌ها، طبقه‌بندی K نزدیک‌ترین همسایه است که این طبقه‌بندی، نمونهٔ آزمون را متعلق به کلاسی می‌داند که بیشترین آرا را در بین k نزدیک‌ترین همسایگان آن داشته باشد. طبقه‌بندی KNN به دلیل قابلیت درک بالا و عدم نیاز به ایجاد فرضیه روی داده‌ها، روشی ساده و پرکاربرد می‌باشد.

۸. روش شبکهٔ عصبی مصنوعی[17]: علی‌رغم به‌کارگیری یک ساختمان ساده در این روش، سرعت و قدرت محاسباتی آن به‌شدت مورد توجه قرار گرفت. روش شبکهٔ عصبی از سیستم عصبی الهام گرفته و از ساختار مغز و اعصاب انسان پیروی می‌کند (ریچارد[18]، ۲۰۱۳). امروزه مدل شبکهٔ عصبی در بخش‌های مختلف علوم به‌منظور مدل‌سازی روابط پیچیدهٔ غیرخطی به‌کاررفته گرفته می‌شود و تا حدودی جایگزین مدل‌های آماری شده است؛ زیرا شبکه‌های عصبی مصنوعی بدون نیاز به حل معادلات دیفرانسیل جزئی، غیرخطی بودن فرایند موردنظر را شبیه‌سازی می‌کنند و حتی زمانی که مجموع داده‌های آموزشی حاوی داده‌های خطادار باشد، عملکرد مناسبی را نشان می‌دهند. شبکه‌های چندلایه دارای توانایی بیشتری بوده که در آن یک لایهٔ ورودی وجود دارد که اطلاعات را دریافت می‌کند و تعدادی لایهٔ مخفی دارد که

1. Pinto
2. LOLIMOT (Local Linear Model Tree)
3. Support vector machines
4. Supervised Learning
5. Optimized
6. Linear
7. Polynomial
8. Radial Basis
9. Hyperbolic Tangent
10. Laplacian
11. Bessel
12. Anova RBF
13. Spline
14. Ouralbena
15. K Nearest Neighbor
16. Nearest Neighbor Analysis
17. Artificial Neural Network
18. Richard

اطلاعـات را از لایـه‌های قبلـی می‌گیـرد و در نهایـت، یـک خروجی وجـود دارد کـه نتیجـهٔ محاسـبات بـه آن‌هـا رفتـه و خروجی آن، خروجی نهایی شبکه است (آبراهام، ۲۰۰۵).

۹. روش رگرسیون لجستیک[1]: رگرسیون لجسـتیک یکـی از انواع مدل‌های خطی تعمیم‌یافته است که برای تحلیل وجـود یا عدم وجود متغیر وابسته بسیار مناسب می‌باشد. ایـن مـدل از تابع لوجیت به‌عنوان تابع پیوند اسـتفاده می‌کنـد و خطـایش از توزیع چندجمله‌ای پیروی می‌کند. هدف از تحلیل رگرسـیونی لجستیک، دستیابی به مدلی مناسب و در عین حال سـاده بـرای بررسی ارتبـاط بـین متغیـر وابسـته بـا یـک یـا مجموعـه‌ای از متغیرهای مستقل است (چن و وانگ[2]، ۲۰۰۷).

۱۰. روش تشخیص الگو[3]: این روش شـاخه‌ای از هـوش مصنوعی است که با طبقه‌بندی و توصیف مشـاهدات سـروکار دارد. شناسایی الگو به ما کمک می‌کنـد داده‌هـا را بـا تکیـه بـر دانـش قبلـی یـا اطلاعـات آمـاری استخراج‌شـده از الگوهـا، طبقه‌بندی نماییم. تشخیص الگو عبارت اسـت از دسـته‌بندی و تفکیـک الگوهـای خـاص بـر اسـاس ویژگی‌هـای از پـیش تعریف‌شده از مجموعه‌ای از داده‌های در دسـترس. ایـن روش به‌اجمال، با مسـائل خوشـه‌بندی و طبقه‌بنـدی سـروکار دارد و دربرگیرندهٔ طیـف گسـترده‌ای از روش‌هـای آمـاری کلاسـیک الگوریتم‌های هوشمند شبکه‌های عصبی و منطـق فـازی اسـت (هان و کامبر[4]، ۲۰۰۱).

۱۱. روش گروهی مدل‌سـازی داده‌هـا[5]: ایـن الگـوریتم قابلیت استفاده در موضـوعات متنـوعی چـون کشـف روابـط، پیش‌بینـی، مدل‌سـازی، بهینه‌سـازی و شـناخت الگوریتم‌هـای غیرخطـی را دارد. ویژگـی خـاص ایـن الگـوریتم اسـتنتاجی، قابلیت شناسایی و غربـال کـردن متغیرهـای کم‌اثـر ورودی در دورهٔ آموزشی شبکه و حذف آن‌ها از روند شبیه‌سازی در دورهٔ آزمون است. بدین ترتیب می‌توان با انجام یک فراینـد قیاسـی، در چند مرحلهٔ تکرار، متغیرهای کم‌اثرتر را حذف کرد و نهایتـاً مدل بهینه و پیش‌بینی را بر اساس معیارهای متـداول خطـا بـه دست آورد (کاکائی لفدانی، ۲۰۱٤).

برازش الگوریتم‌ها

همان‌طور که گفته شد، برنامه‌نویسی هریک از این الگوریتم‌هـا در نرم‌افزار MATLAB صورت گرفت. بدین صـورت کـه در تمامی الگوریتم‌ها، تعداد ۷۰درصد داده‌ها بـه‌منظور آمـوزش و ۳۰درصد باقی‌مانده برای آزمون (صحت‌سنجی) انتخاب شـدند و خروجی‌هـای مـورد نظـر اسـتخراج گردیـد تـا بهتـرین الگوریتم‌ها در تفکیـک منـابع رسـوبی حوضـه (هفـت واحـد سنگ‌شناسی به‌عنوان متغیر هدف)، در مراحل آموزش و آزمون تعیین گردد.

ارزیابی الگوریتم‌ها بر اساس چهار گـروه ورودی مختلـف اجرا شد. مرحلهٔ اول بر اساس ۲۱ متغیرهای ژئوشیمیایی شامل عناصر B, Ce, Co, Cr, Cu, K, Li, Na, P, Pb, Ti, V, Y, Zn, Ba, Mn, Sr, Mg, Al, Ca و Fe ، مرحلـهٔ دوم بـر اسـاس ۱۱ گروه سنگ‌شناسی همگن شامل کوارتز، توف، لاتیت، داسـیت، آندزیت، دولومیت، کلسیت، توف آنـدزیتی، آنـدزیت‌لیتیک‌دار، پیروژنیک و نمک با استفاده از نمونه‌های خرده‌سنگ‌ها و ذرات بـا قطـر بـیش از ٤۰۰۰، ۲۰۰۰، ۱۰۰۰، ۵۰۰، ۲۵۰، ۱۲۵، ٦۳ و کوچک‌تر از ٦۳ میکرون و مرحلـهٔ سـوم بـر اسـاس دانه‌بنـدی ذرات شامل پارامترهای D10, D50, D90، درصد شن، درصـد سیلت، درصد رس، چولگی و کشیدگی ذرات انجـام شـد. در مرحلـهٔ چهـارم نیـز بـا ورود تمـامی متغیرهـای ژئوشـیمیایی، دانه‌بندی و سنگ‌شناسی، الگوریتم‌های طبقه‌بندی اجرا شد.

در انتهـا بـرای حصـول اطمینـان از رونـد مدل‌سـازی، اعتبارسـنجی و ارزیـابی دقـت مـدل‌های مـورد اسـتفاده، از معیارهای آماری ضـریب تبیـین[6] (R^2) و میـانگین مربـع خطـا[7] (MSE) استفاده شد که به‌ترتیب از روابط (۱) و (۲) بـه دسـت می‌آیند که در این روابط مؤلفه n، تعداد داده‌هـای مورد ارزیابی و x_i و y_i، امین داده برآوردی و اندازه‌گیری شده است.

$$R^2 = \sum (y_i - x_i)^2 / \sum (y_i - x_i)^2 \quad (۱)$$

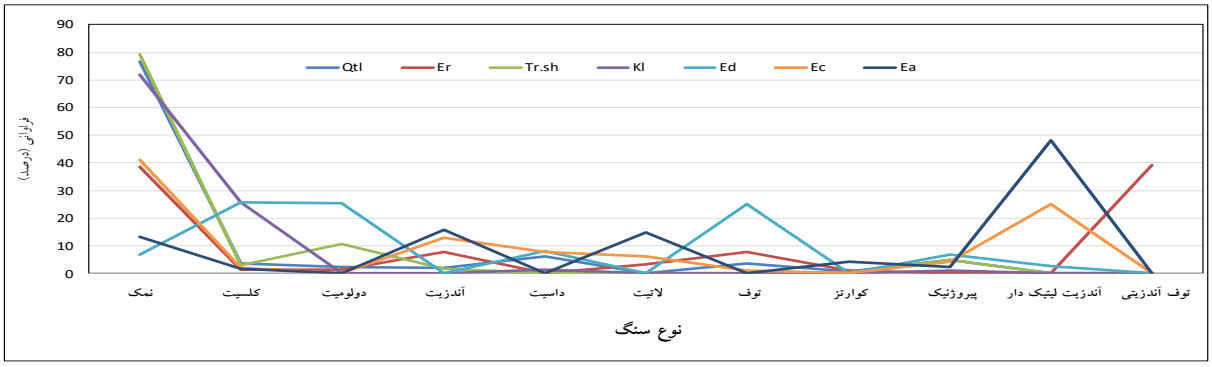
$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2 \quad (۲)$$

1. Logistic Regression
2. Chen and Wang
3. Pattern Recognition
4. Han and Kamber
5. Group Method of Data Handling
6. Coefficient of Determination
7. Mean Squared Error

نتایج

نمودار متوسط مقادیر فراوانی نمونه سنگ‌ها، متغیرهای دانه‌بندی و متوسط غلظت عناصر ژئوشیمیایی در شکل (۲) و جداول (۱) و (۲) نمایش داده شده است. در بحث سنگ‌شناسی سنگریزه‌های قشر بالایی رسوبات پشت مخزن سد بیشتر از جنس کلسیت و نمک بوده، در رسوبات آبرفتی قدیم و گدازه‌های آتشفشانی بیشترین نوسانات فراوانی سنگ‌ها را داریم. در واحد نهشته‌های آبرفتی قدیم نیز پس از نمک که بیشترین مقدار بوده، سایر سنگ‌ها به‌دلیل خاصیت رسی و سیلتی کاهش یافته تا مواردی مانند پیروژنیک، آندزیت لیتیک‌دار تقریباً به صفر رسیده است. به این دلیل که نهشته‌های قدیمی اغلب کنگلومرایی هستند و مقادیر رسی و سیلتی کمی را دارند. کاهش یافته مقادیر کلسیت، دولومیت و توف مقدار زیادی را به خود اختصاص داده است؛ زیرا در نهشته‌های قدیم کمتر رسی و سیلتی وجود دارند و اغلب کنگلومرایی می‌باشند. بنابراین مقدار آندزیت لیتیک دارو توف آندزیتی در آن‌ها کم است و در نهشته‌های آبرفتی حاضر به مقدار زیاد وجود دارند. جدول (۱) نشان می‌دهد تا D50، متغیر دانه‌بندی روند ثابتی داشته و سپس به‌صورت افزایشی عمل کرده و D90، بیشترین مقدار غلظت را در واحدهای زمین‌شناسی نشان می‌دهد به این دلیل که قطری است که ۹۰درصد از ذرات از آن کوچک‌ترند و مجدد روند نزولی را ارائه کرده است. نتایج مربوط به بررسی غلظت عناصر ژئوشیمیایی نشان داد غلظت عناصرژئوشیمیایی در هفت گروه زمین‌شناسی مورد بررسی از میزان متوسط تا زیاد برخوردار است. به‌طور کلی به‌ترتیب عناصر Ca، Fe، Mg، AL دارای بیشترین غلظت در نمونه‌های خاک بودند و سایر عناصر به میزان کمتر وجود داشتند. به‌ویژه B و Co کمترین مقدار را به خود اختصاص دادند. در بین واحدهای زمین‌شناسی، آهک اوربیتولین‌دار بیشترین سهم را در رسوبات مخزن سد دارد.

همچنین نتایج مربوط به طبقه‌بندی منابع رسوبی پس از ورود داده‌های آزمایشگاهی مورد نظر به محیط نرم‌افزار MATLAB و برازش الگوریتم‌ها، در قالب جدول‌های (۳) تا (٦) در مراحل آموزش و آزمون ارائه شده‌اند.



شکل (۲): متوسط فراوانی سنگ‌ها در واحدهای مختلف زمین‌شناسی حوضه با تحلیل مورفوسکوپی نمونه‌ها

Figure (2): The average frequency of rocks in different geologic units of watershed using morphoscopic analysis of samples

جدول (۱): متوسط متغیرهای دانه‌بندی رسوبات در واحدهای مختلف زمین‌شناسی حوضه

Tabel (1): The average of granulometric variables of sediments in different geologic units of watershed

زمین‌شناسی	واحد	درصد رس	درصد سیلت	درصد شن	٤< mm	۲< mm	۱< mm	٥۰۰< µm	۲٥۰< µm	۱۲٥< µm	٦۳< µm	٦۳> µm	D10	D50	D90	چولگی	کشیدگی
Qtl		۲۰/٥	٤۰/۹	۳۸/٥	۲۷/۸	٦/٤۱	۷/۱	٤/۸	۳/٤	٦/۷	٤۷/۳	۸٥/۰	۷/۷	٦۱/۷	۳۳۳٦/۱	۰/۷	۲/٤
Er		۱۹/٥	۲۲/۷	٥۷/۸	۳۲/۸	۲۷/۲	۳۳/۱	۱۸/۳	۸/۲	٦/۸	۱۷/٦	٥۰/٦	۱۲/۷	۷۸۲/۹	٤٥۰٥/٦	-۰/٥	۱/۹
Tr.sh		۲۳/٦	٤۲/٦	۳۳/۸	۳۳/۰	۱۰/۰	٦/۷	٤/۸	۷/۰	۷/۸	۳۹/۰	۱٦٦/۰	٦/۲	۳۹/۳	٤۲۲۳/۱	۱/۰	۲/٦
Kl		۱۸/۱	٤۹/۸	۳۲/۱	٦۰/۷	۱٤/۰	٦/۲	۲/۳	۱/۹	۳/٦	۳۲/۰	۷۲/٦	۸/۸	۷٦۸/٤	٤۹۷۹/۷	۰/۱	۱/٦
Ed		۱۰/٥	۲۰/٦	٦۸/۹	۱/۳	۳/۹	۲۱/٤	۲۳/۸	۳۳/۷	۳٤/۰	٤۷/۱	۲۸/۳	۲۸/۲	۱٥٦/٥	۱۱۱۸/۹	-۰/٥	۲/۷
Ec		۱۰/۳	۳۱/۹	٥۷/۸	٦۱/٤	۱٦/۸	۱٦/۷	۸/٦	۷/۱	۱۱/٤	۳۳/٦	٤۱/۲	۲۷/٦	۸۹۳/۱	٤۹۲٥/٥	-۰/٤	۲/٤
Ea		۹/۰	۱۱/۳	۷۹/۸	۲۰/۷	۱۰/۳	۲٦/۹	۳۰/٤	۳۳/۸	۲۲/٤	۲۱/۸	۲٥/۸	۳۹/٥	٤۰٤/٦	۳۱۸۲/۱	-۰/٦	۲/۹

جدول (۲): متوسط غلظت عناصر ژئوشیمیایی در واحدهای مختلف زمین‌شناسی حوضه

Tabel (2): The average concentration of geochemical elements sediments in different geologic units of watershed

واحد زمین‌شناسی	B	Ce	Co	Cr	Cu	k	Li	Na	P	Pb	Ti	V	Y	Zn	Ba	Mn	Sr	Mg	Al	Ca	Fe
Qtl	۰/۰۸	۰/۲۲	۰/۲۲	۰/۷۱	۰/۵۵	٤/٦۰	۰/۵۸	۲۸/۰۵	٦/۲۵	۰/۲۲	۰/۵۷	۰/٤٤	۰/۱۹	۱/۳٤	۲/۷۳	۱۰/۹٤	۳/۳٤	۳٦۲/۲۵	۱۳٤/۳۷	۱۱٤۸/۷۱	٤۰۵/۸۲
Er	۰/۰٦	۰/۸٦	۰/۱۸	۰/۵۰	۰/٤۹	٤/۸۲	۰/٤۳	۲۲/۸۹	۳/٦۷	۰/۲۸	۰/۵٦	۰/۲۸	۰/۲۷	۱/۱٦	٤/۷۱	۹/٤۱	۲/۰۲	۲۸٦/٤٤	۹٦/٦۷	٦۲۲/۸۷	۲۷۳/٦۱
Tr.sh	۰/۰٦	۰/۲۱	۰/۲۲	۰/۷۱	۰/۵۷	٤/۸۷	۰/٦۰	۲۷/٤٤	٤/۹۱	۰/۱۸	۰/٤۷	۰/٤٤	۰/۱۸	۱/۲۷	۲/۸۲	۹/۳۳	۵/۱٦	۳٤۲/۰۲	۱۲۸/۹۲	۱۸۸٤/٦۰	٤۲۳/٤۹
Kl	۰/۰۸	۰/۱۷	۰/۲۵	۰/۷٦	۰/۵۵	۳/۹۳	۰/۵۹	۳۱/٦۲	۵/٤۹	۰/۲۵	۰/٦٦	۰/٤۹	۰/۱۸	۱/۳۸	۲/٤٦	۱۱/۸۳	۳/۱۷	٤۲۹/۹۷	۱۰۹/٦۳	۱۳۷٤/۰۸	٤۲۸/۵٤
Ed	۰/۰۸	۰/۲۸	۰/۱٦	۰/۲۵	۰/۳۱	۲/٤۰	۰/۳۹	۲۵/۷۳	۱/٦۱	۰/۱۲	۰/۳۲	۰/۱۷	۰/۱٤	۱/۰۱	۲/٤۳	۸/۸٤	۲/۳٤	٤۰۳/۵٤	۵۱/٤٦	۱۳۳٤/۲۰	۱۲۱/٦٤
Ec	۰/۰٦	۰/٤۸	۰/۱۲	۰/۵۵	۰/۲٦	۲/۸٤	۰/۳٤	۲۰/٤۱	۵/۳۸	۰/۲۵	۰/۲۳	۰/۳۷	۰/۱۵	۰/۷۱	۱/۷٤	۷/٤۹	۱/۷۱	۲۵۸/۲۱	۷۹/۳۲	٦۳۹/٦۷	۲۰٤/۷۷
Ea	۰/۰۷	۰/٤۸	۰/۱۸	۰/٦٦	۰/٤۳	۳/۰٤	۰/۵۰	۳٤/۰۰	۵/۹۰	۰/۲٦	۱/۳۳	۰/۵۰	۰/۲٤	۱/۲۳	۲/۰۸	۱۰/۹۸	۲/۰۷	۲۹۹/۱۱	۹۷/۱٦	٦٤۲/۷۵	٤٤۷/۲۲

جدول (۳): نتایج طبقه‌بندی منابع رسوبی با ورود متغیرهای ژئوشیمیایی

Tabel (3): The results of sediment sources classification using geochemical variables

نوع الگوریتم	ضریب تبیین		میانگین مربع خطا	
	(آموزش)	(آزمون)	(آموزش)	(آزمون)
رگرسیون خطی	۰/٤۹	۰/۱۹	۰/٦۱	۷/۸٦
رگرسیون درجه‌دوم خالص	۰/۵۳	۰/۲۸	۰/٤۷	۳/٤۸
رگرسیون اثر متقابل	۰/۵۰	۰/۲٤	۰/۵۰	٤/۷۱
رگرسیون درجه‌دوم	۰/٦۲	۰/۱٤	۰/۳۸	۵/۷۱
شبکهٔ عصبی	۰/۳۱	۰/۱٤	۲/۱۱	۳/۱۳
روش تشخیص الگو	۰/۲۷	۰/۱٦	۲/۱٦	۳/۲۲
تحلیل ممیزی	۱	۰/٤۸	۰	۸/۰۵
درخت طبقه‌بندی	۰/۷۰	۰/۲٤	۱/۷۰	۵/۱٤
درخت رگرسیونی	۰/۳۸	۰/۲۸	۰/۷۹	۲/۰۰
جنگل تصادفی	۱	۰/٤۰	۰	۳/۷۸
نزدیک‌ترین همسایه	۱	۰/۵۲	۰	۲/۷٦
ماشین بردارپشتیبان خطی	۱	۰/٦۲	۰	۳/۰۵
ماشین بردارپشتیبان چندجمله‌ای	۱	۰/٤۳	۰	٤/٦۵
ماشین بردارپشتیبان لاپلاسین	۱	۰/٦۷	۰	۱/٤۸
ماشین بردارپشتیبان تابع شعاع مبنا	۰/٦۸	۰	۱/۳۸	۳/۹۵
ماشین بردارپشتیبان چندگانه	۱	۰/۳۳	۰	۵/۸٦
روش گروهی مدل‌سازی داده‌ها	۰/٦۷	۰/۲۲	۰/٤٦	۵/۲۱

جدول (٤): نتایج طبقه‌بندی منابع رسوبی با ورود متغیرهای دانه‌بندی

Tabel (4): The results of sediment sources classification using granulometric variables

نوع الگوریتم	ضریب تبیین		میانگین مربع خطا	
	(آموزش)	(آزمون)	(آموزش)	(آزمون)
رگرسیون خطی	۰/٤٦	۰/۱٤	۰/۷۱	۷/۷٦
رگرسیون درجه دوم خالص	۰/٦۲	۰/۲۸	۰/۳۸	٦/۱۰
رگرسیون اثر متقابل	۰/۵٤	۰/۱٤	۰/٤٦	۹/۵۷
رگرسیون درجه دوم	۰/۵٤	۰/۱۹	۰/٤٦	۳/٤۸
شبکهٔ عصبی	۰/۲۷	۰/۱۳	۲/۷۱	٤/۲۱
روش تشخیص الگو	۰/۲٤	۰/۱۹	۳/۳۱	٤/۱۷
تحلیل ممیزی	۱	۰/۱٤	۰	۹/٦۲
درخت طبقه‌بندی	۰/۷۵	۰	۰/۹۰	٦/۲۸
درخت رگرسیونی	۰/۳۵	۰/۱٤	۰/۹٤	۲/۸٦
جنگل تصادفی	۱	۰/٤۳	۰	٦/۱۰
نزدیک‌ترین همسایه	۱	۰/۳۸	۰	۵/۹۰
ماشین بردارپشتیبان خطی	۱	۰/٤۸	۰	۵/۲۸
ماشین بردارپشتیبان چندجمله‌ای	۱	۰/۵۱	۰	٤/۹۵
ماشین بردارپشتیبان لاپلاسین	۱	۰/٤۸	۰	۵/۵۲
ماشین بردارپشتیبان تابع شعاع مبنا	۰/٦٦	۰	۱/٦۲	٦/۲٤
ماشین بردارپشتیبان چندگانه	۱	۰/٤۳	۰	٤/۸۱
روش گروهی مدل‌سازی داده‌ها	۰/٦۵	۰/۲۷	۰/۵۱	٤/۹۸

جدول (۵): نتایج طبقه‌بندی منابع رسوبی با ورود متغیرهای سنگ‌شناسی

Tabel (5): The results of sediment sources classification using geologic variables

نوع الگوریتم	ضریب تبیین		میانگین مربع خطا	
	(آموزش)	(آزمون)	(آموزش)	(آزمون)
رگرسیون خطی	۰/۳۹	۰/۲٤	۱/۰٤	۵/۰۵
رگرسیون درجه‌دوم خالص	۰/٦۵	۰/۱٤	۰/۳۵	۹/۷٦
رگرسیون اثر متقابل	۰/٦۵	۰/۱۹	۰/۳۵	٤/۵۲
رگرسیون درجه‌دوم	۰/٦۱	۰/۲۸	۰/۳۹	٤/۸۱
شبکهٔ عصبی	۰/۳	۰/۳۰	۲/۵۰	٤/۸۷
روش تشخیص الگو	۰/۲٤	۰/۱۳	۳/٤۰	۵/۰۸
تحلیل ممیزی	۱	۰/٦۳	۰	۵/۱۰
درخت طبقه‌بندی	۰/۷۵	۰	٤/۳۲	۱٤/۸۱
درخت رگرسیونی	۰/۳٤	۰/۱۹	۱/۱۷	٦/۳۳
جنگل تصادفی	۱	۰/۷۵	۰	۱/۹۲
نزدیک‌ترین همسایه	۱	۰/٦۲	۰	۳/۷۱
ماشین بردارپشتیبان خطی	۱	۰/٦۲	۰	۲/۰۵
ماشین بردارپشتیبان چندجمله‌ای	۱	۰/٦۰	۰	۲/۳۰
ماشین بردارپشتیبان لاپلاسین	۱	۰/٦۲	۰	۲/۱۹
ماشین بردارپشتیبان تابع شعاع مبنا	۰/۸۷	۰/۲۸	۱/۳٦	٤/۳۳
ماشین بردارپشتیبان چندگانه	۱	۰/٦۲	۰	۵/۸۱
روش گروهی مدل‌سازی داده‌ها	۰/٦٦	۰/٤۱	۰/٤۸	۷/۳۳

جدول (٦): نتایج طبقه‌بندی منابع رسوبی با ورود تمامی متغیرها

Tabel (6): The results of sediment sources classification using whole variables

نوع الگوریتم	ضریب تبیین		میانگین مربع خطا	
	(آموزش)	(آزمون)	(آموزش)	(آزمون)
رگرسیون خطی	۰/۵۸	**۰/۰۹**	۰/٤۱	۱۰/۱۹
رگرسیون درجه‌دوم خالص	۰/۵۵	**۰/۲۳**	۰/٤٤	٦/٤۲
رگرسیون اثر متقابل	۰/۵۱	**۰/۱٤**	۰/٤۸	۹/٤۷
رگرسیون درجه‌دوم	۰/۵٦	**۰/۰۹**	۰/٤۳	۸/٦۱
شبکهٔ عصبی	۰/۳۳	**۰/۲۳**	۲/۲٦	۳/۵۳
روش تشخیص الگو	۰/۲۸	**۰/۱۵**	۲/۷۳	٤/٤٤
تحلیل ممیزی	۱	**۰/۷٦**	۰	۲/۹۰
درخت طبقه‌بندی	۰/۷۵	۰	۳/٤۷	۱٤/٦٦
درخت رگرسیونی	۰/۲۵	**۰/۱٤**	۰/۹۱	٤/٤۷
جنگل تصادفی	۱	**۰/٦۰**	۰	۲/۸۵
نزدیک‌ترین همسایه	۱	**۰/۷۱**	۰	۱/۱۹
ماشین بردارپشتیبان خطی	۱	**۰/۷۱**	۰	۳/۸۵
ماشین بردارپشتیبان چندجمله‌ای	۱	**۰/۵۵**	۰	۳/٤۲
ماشین بردارپشتیبان لاپلاسین	۰/۹۵	**۰/٤۲**	۰/۱۳	۲/۸۰
ماشین بردارپشتیبان تابع شعاع مبنا	۱	**۰/۷٦**	۰	۱/۵۲
ماشین بردارپشتیبان چندگانه	۱	**۰/٤۷**	۰	٦/٦۱
روش گروهی مدل‌سازی داده‌ها	۰/٦۸	**۰/۲٦**	۰/٤۰	٦/۷۱

بحث و نتیجه‌گیری

برای تعیین الگوریتم‌های مناسب جهت تفکیک منابع رسوبی حوضهٔ مطالعاتی، از مقایسهٔ مقدار ضریب تبیین و میانگین مربع خطا به‌صورت جداگانه برای داده‌های آموزش و آزمون استفاده شد. به این صورت که هرچه مقدار ضریب تبیین بیشتر و میانگین مربع خطا کمتر باشد، نشان‌دهندهٔ دقت بیشتر و در نتیجه عملکرد بهتر الگوریتم است. طبق نتایج ارائه‌شده، به‌طور کلی دقت الگوریتم‌های طبقه‌بندی در مرحلهٔ آزمون بهتر از مرحلهٔ آزمون بوده و خطای کمتری را نیز به همراه داشته است. نکتهٔ قابل ذکر اینکه در اجرای الگوریتم‌ها با چهار نوع ورودی ژئوشیمیایی، سنگ‌شناسی، دانه‌بندی و ورود توأم سه گروه متغیر، ماشین‌های بردارپشتیبان بالاترین دقت و کمترین خطا را در طبقه‌بندی منابع رسوبی داشته‌اند. بنابراین با توجه به دقت مناسب روش‌های به‌کاررفته و ارائهٔ نتایج موفقیت‌آمیز مدل‌ها برای تفکیک منابع رسوبی حوضه بهتر است تمامی عوامل فوق دخالت داده شوند. طبق جداول فوق بیشترین دقت طبقه‌بندی منابع رسوبی از ورود متغیرهای ژئوشیمیایی به‌دست آمده و پس از آن ورود توأم سه دسته متغیر، بیشترین دقت را داشته است. متغیرهای سنگ‌شناسی و دانه‌بندی نیز به‌ترتیب منجر به کمترین دقت طبقه‌بندی توسط الگوریتم‌ها شده‌اند.

نتایج تفسیر و مقایسهٔ خروجی‌های الگوریتم‌های به‌کاررفته نشان داد که در مرحلهٔ آموزش، الگوریتم‌های تحلیل ممیزی، جنگل تصادفی، k نزدیک‌ترین همسایه، ماشین بردارپشتیبان خطی[1]، ماشین بردارپشتیبان چندجمله‌ای[2]، ماشین بردار چندگانه[3] و ماشین بردار با تابع شعاع مبنا[4] با حداکثر مقدار ضریب تبیین (R^2=۱) و مقدار میانگین مربع خطا برابر صفر، بهترین عملکرد را در تفکیک منابع رسوبی حوضه نشان دادند و به‌عنوان بهترین الگوریتم‌های پژوهش معرفی شدند. لذا الگوریتم‌های به‌کاررفته در پژوهش در تفکیک منابع رسوبی دقت بیشتری را از خود نشان دادند که با نتایج حدادچی و همکاران (۲۰۱۳) در انگشت‌نگاری رسوبات آبرفتی همخوانی

1. Linear Support Vector Machine
2. Polynominal Support Vector Machine
3. Multiple Support Vector Machine
4. Radial Basis Function (RBF)

دارد. اما مغایر با مطالعات پیشین، الگوریتم شبکهٔ عصبی نسبت به سایر الگوریتم‌ها عملکرد بهتری را در تفکیک منابع رسوبی حوضه نشان نداد. همچنین روش درخت رگرسیونی با مقادیر ضریب تبیین ۰/۲۸ و خطای ۰/۹۱، دارای ضعیف‌ترین عملکرد در این زمینه است. در میان داده‌های آزمون نیز بیشترین مقدار ضریب تبیین به دو الگوریتم تحلیل ممیزی و ماشین بردارپشتیبان با تابع شعاع مبنا با مقدار برابر ۰/۷۶ اختصاص‌یافته که با مقایسهٔ مقادیر خطای آن‌ها مشخص می‌شود الگوریتم ماشین بردار با تابع شعاع مبنا (RBF) با میزان خطای کمتر، دقیق‌ترین عملکرد را ارائه کرد و ضعیف‌ترین عملکرد مربوط به الگوریتم درخت طبقه‌بندی است.

همچنین در مرحلهٔ آزمون نیز، دقیق‌ترین الگوریتم باز هم مربوط به خانوادهٔ ماشین‌های بردارپشتیبان با هستهٔ تابع شعاع مبنا بوده که با مطالعهٔ میسرا و همکاران (۲۰۰۹) در شبیه‌سازی رواناب و رسوب همخوانی دارد. همچنین الگوریتم درخت طبقه‌بندی با بیشترین میزان خطا، ناکارآمدترین روش در این مرحله معرفی شده است. اما در پژوهش ستاری و همکاران (۲۰۱٦) بر روی رسوبات معلق رودخانه‌ای و کومارگویال (۲۰۱٤) بر روی داده‌های تولید رسوب، الگوریتم درخت M5 به‌عنوان روش برتر گزارش شد. بررسی پژوهش‌های انجام‌شده در زمینهٔ فرسایش و رسوب می‌توان نتیجه گرفت الگوریتم شبکهٔ عصبی نسبت به سایر الگوریتم‌ها نتایج دقیق‌تری را ارائه کرده و الگوریتم برتر معرفی شده است؛ از جمله مطالعهٔ حرما و همکاران (۲۰۱۵) در مطالعهٔ داده‌های رواناب و رسوب، ملسی و همکاران (۲۰۱۱) در بررسی رسوبات معلق رودخانه‌ای، زو و همکاران (۲۰۰۷) در مطالعهٔ رسوبات معلق و نائینی و همکاران (۲۰۰۸) در بررسی داده‌های غلظت رسوب. اما در پژوهش حاضر این موضوع تأیید نشد و روش شبکهٔ عصبی کمترین دقت و ضعیف‌ترین عملکرد را ارائه نمود که می‌تواند ناشی از تفاوت در نوع داده‌های ورودی و همچنین استفاده از ۱۰ الگوریتم برتر داده‌کاوی در کنار شبکهٔ عصبی مصنوعی باشد. همچنین در مطالعه‌ای که توسط کازوگلو و سالکسن[1] در سال ۲۰۰۹ برای طبقه‌بندی پوشش سطحی اراضی انجام گرفت، دلیل دقت بالای الگوریتم SVM نسبت به دیگر روش‌های داده‌کاوی این‌طور بیان شد که این روش ازلحاظ محاسباتی بسیار سریع بوده و از قوانین بهینه‌سازی برای مکان‌یابی مرزهای بهینه بین کلاسه‌های کاربری استفاده می‌کند. در نتیجه می‌توان آن‌ها را به‌عنوان جایگزین مناسبی برای سایر الگوریتم‌های طبقه‌بندی نیز معرفی کرد. البته باید اذعان نمود برای همهٔ مسائل الگوی تعیین‌شده‌ای وجود ندارد و تعیین بهترین الگو برای هر مسئله نیازمند انجام سعی و خطاهای مکرر و آزمون روش‌ها در شرایط مد نظر است. بنابراین توصیه می‌شود در مطالعات دیگر نیز قابلیت‌های این موارد مورد بررسی قرار گیرد تا نتایج دقیق‌تری به دست آید. در مجموع می‌توان گفت پژوهش‌های انجام‌شده در سال‌های اخیر مانند حیات‌زاده و همکاران (۲۰۱۵)، کاکائی لفدانی (۲۰۱۳)، اولکی[2] و همکاران (۲۰۱٦) و سایر پژوهشگران حاکی از برتری روش‌های هوش مصنوعی نسبت به سایر روش‌ها در محاسبات مربوط به فرسایش و رسوب است. پس می‌توان اذعان کرد داده‌کاوی روشی تحلیلی است که امروزه به‌طور وسیعی در تحقیقات پژوهشی کاربرد ویژه‌ای دارد و با گسترش و تنوع روزافزون الگوریتم‌ها، لزوم تحقیق و بررسی در زمینه‌های گوناگون مدل‌ها را ایجاب می‌کند. بنابراین استفاده از روش الگوریتم‌های طبقه‌بندی با توجه به محاسبات دقیق و عدم نیاز به‌صرف هزینه و وقت زیاد و ابزارهای پیشرفته توصیه می‌گردد.

1. Kavzoglu and Colkesen

2. Ulke

منابع

1. Abraham, A., 2005. Artificial neural networks, Oklahoma State University, Stillwater, USA. 908 PP.
2. Arabkhedri, M., 2014. An overview of the effective factors on the water erosion in Iran. Journal of Land Management, 2 (1): 23-35.
3. American Society for Testing and Materials (ASTM), 2008. Standard test method for particle-size analysis of soils. In: Annual Book of ASTM Standards. Philadelphia.
4. Besalatpour, A.A., Ayoubi, S.A., Hajabasi, D.A., 2016. Gamma test to select the optimal inputs in modeling soil shear strength using artificial neural network. Journal of Soil and Water Conservation Researches (Agricultural Sciences and Natural Resources). 20 (1): 97-114.
5. Breiman, L., 2001. Application and analysis of random forests and machine learning. Journal of Water Management, 15(1): 5-32.
6. Breiman, L., Friedman J., Olshen R., and Stone, C., 1984. Classification and Regression Trees, Chapman & Hall/CRC Press, Boca Raton, FL.
7. Chan, C., Lewis, B., 2002. A basic primer on data mining, Information Systems Management. Journal information System Management, 19(4): 56-69.
8. Chen, Zh. And Wang, J., 2007. Landslide hazard mapping using logistic regression model in MackenzieValley, Canada. Geomorphology, Vol.42.
9. Demirci, M., Baltaci, A., 2013. Prediction of suspended sediment in river using fuzzy logic and multi linear regression approaches. Neural Computing and Applications, 23 (1): 145-151.
10. Feyznia, S., 2008. Applied sedimentology with emphasis on soil erosion and sediment production. Gorgan University of agricultural sciences and natural resources press, 356 pp.
11. Haddadchi, A., Ryder, D.S., Evrard, O. and Olley, J., 2013. Sediment fingerprinting in fluvial systems: review of tracers, sediment sources and mixing models. International Journal of Sediment Research, 28(4): 560-578.
12. Harma, N., Zakaullah, M.D., Tiwari, H., Kumar, D., 2015. Runoff and sediment yield modeling using ANN, and support vector machines (case study: from Nepal watershed). Ore Geology Reviews, 17(9): 63-89.
13. Han, D. and Kamber, M., 2001. Data Mining: Concepts and Techniques. San Diego Academic Press.
14. Hayatzadeh, m., Chezgi, G., Dastorani, M.T., 2015. Evaluation of sediment rating curve and neural network using a combination of morphological parameters Baghabas area. Journal of Agricultural Sciences and Natural Resources, 19 (70): 101-119.
15. Joudi, A. R. & Sattari, M. T., 2017. Evaluation of the performance of quenel based methods in estimating the suspended rainfall of river (case Study: Sufy Chay, Maragheh). Journal of Research in Natural Geography Vol 38(33): 413-429.
16. Kakaei-Lafdani, E., Moghaddamnia, A., Ahmadi, A. Ebrahimi, C., 2013. Daily suspended sediment load prediction using artificial neural networks and support vector machines. Journal of Hydrology, 478(25): 50-62.
17. Kavzoglu, T. and Colkesen. I., 2009. A kernel function analysis for support vector machines for land cover classification. Journal of applied Earth Obzervation and Geoinformation, 11(5):352-359.
18. Kakaei Lafdani, E., Pournemat Roudsari, A., Qaderi, K. and Moghaddam-Nia, A., 2014. Predicting the Volume of Suspended Sediments using GMDH and SVM Models Based on Principal Component Analysis. 9th International River Engineering Conference Shahid Chamran University, Ahwaz, pp: 22-24.
19. Keshavarz-Emami, R., karolouks, A., 2007. Local linear tree algorithm development (LOLIMOT) using a fuzzy validity function and credit for time series prediction. 1st Joint Congress on Fuzzy and Intelligent Systems Ferdowsi University of Mashhad, Iran, 29-31 Aug.
20. Kumar Goyal, M., 2014. Modeling of Sediment Yield prediction Using M5 Model Tree Algorithm and Wavelet Regression Journal of Water Resources Management, 28, 1991-2003.
21. Melesse, A. M., Ahmad, M. E., Mcclain, X., Wang, F. and Lim, Y.H., 2011. Suspended sediment load prediction of river systems: An artificial neural network approach. Journal of Agricultural Water Management, 98(5): 855-866.
22. Milhouse, R.T., 1998. Modeling of instream flow needs: the link between sediment and aquatic habitat Soil Sciences. Yazd University publication, Yazd, Iran, 516 pp.
23. Misra, D., Oommenb, T., Agarwal, A., Mishra, A. and Thompson, M., 2009. Application and analysis of support vector machine based simulation for runoff and sediment yield. Biosystems engineering, 6(2): 527- 535.
24. Naeini, S.T., Montazeri, M., Zamani, M.M. and Soltani, F., 2008. Sensitivity analysis of stimulus function of artificial neural network

model in estimating sediment concentration. Proceeding of the 4th National congress of Civil engineering, 17-19 May, Tehran.

25. Oralbona, C., Castellini, B., Caputo, L. and sandini, G., 2010. On-line independent support vector machines pattern Recognition Application. Journal of the International Society for the Prevention and Mitigation of Natural Hazard, 10(6): 127-152.

26. Pinto, U., Maheshwar, B., Shrestha, S. and Morris, C., 2012. Modeling eutrophication and microbial risks in peri-urban river systems using discriminant function analysis, Journal of water research, 46(21): 6476- 6488.

27. Richards, J.A., 2013. Remote Sensing digital image analysis, fifth edition, Springer, 494 pp.

28. Rhoton, F.E., Emmerich, W.E., Nearing, M.A., Mc Chesney, D.S. and Ritchie, J.C., 2011. Sediment source identification in a semiarid Watershed at soil mapping unit scales. Catena, 87: 12-181.

29. Siegel. F.R., 2002, Environmental geochemistry of potentially toxic metals. Springer. Berlin Heidelberg New York, 212 pp.

30. Sattari, M.T., Rezazadehjudi, A., Safdari, F., Ghahramanzadeh, F., 2016. Performance evaluation methods, support vector regression modeling M5 model tree and suspended sediment Ahar Chai River. Journal of Soil and Water Conservation, 6 (1): 109-124.

31. Turnbull, L., Wainwright, J. and Brazier, R. E., 2008. A conceptual framework for understanding semi-arid land degradation: Eco hydrological interaction across multiple-space and time scales. Journal of Ecohydrology, 1(1): 23-34.

32. Ulke, A.G., Tayfur, R., Ozkul, S., 2009. Predicting suspended sediment loads and missing data for Gediz River, Turkey. Journal of Hydrologic Engineering, 14(9): 954-965.

33. Yap, C.A., Esmaeili, A., Tan, S., Omar, H., 2002. Correlations between speciation of Cd, Pb and Zn in sediment and their concentrations in total soft tissue of green- lipped mussel Pernaviridis from the west coast of Peninsular Malaysia. Environment International, 28(1-2): 117-126.

34. Zhu, Y. M., Lu, X. X. and Zhou, Y., 2007. Suspended sediment flux modeling with artificial neural network: An example of the Longchuanjiang River in the Upper Yangtze Catchment. China. Journal of Geomorphology, 84(1): 111-125.

Using Data Mining Algorithms in Separation of Sediment Sources in Nodeh Watershed, Gonabad

Mehdi bashiri1, Mahsa Ariapour[2], Ali Golkarian[3]

Received: 5/03/2018 Accepted: 2/07/2018

Extended Abstract

Introduction: Reduction of sediment supply requires the implementation of soil conservation and sediment control programs in the form of watershed management plans. Sediment control programs require identifying the relative importance of sediment sources, their quantitative *ascription* and identification of critical areas within the watersheds. The sediment source ascription is involves two main steps so that in the first, several diagnostic tracers are selected for obvious and significant separation of potential sources of sediment and in the second step selected tracers for potential sources of sediment are compared, with corresponding values extracted from the sediment samples taken in the watershed outlet. Also, due to the large amount and complexity of data available, nowadays in geo- and environmental sciences, we face the need to develop and incorporate more robust and efficient methods for their analysis and modelling. Therefore recent fundamental progress in data mining algorithms can considerably contribute to the development of the emerging field - environmental data science.

Methodology: According to what was said, in this research, the data mining algorithms used to separate sediment sources in the Nodeh watershed of Gonabad located in Razavi-Khorasan province by using the geochemical (includes the 21 elements of Mg, Sr, Mn, Ba, Zn, Y, V, Ti, Pb, P, Na, Li, K, Cu, Cr, Co, Ce, B, Ca, Al and Fe), granulometric (includes the D_{90}, D_{50}, D_{10}, percent of sand, percent of silt, percent of clay, skewness and kurtosis and the diameters less than 1, 2 and 4 millimeters and less than 500, 250, 125 and 63 microns) and lithological variables (includes the quartz, tuff, laterite, dacite, andesite, dolomite, calcite, andesitic tuff, lithic andesite and salt*)*. A set of 11 classification algorithms includes the decision tree, random forest, regression methods, discriminant analysis, local linear model tree, nearest neighbor analysis, support vector machine, logistic regression, artificial neural network, pattern recognition and group method of data handling programmed in the MATLAB software and the results compared based on the coefficient of determination and mean squared error.

Results and Discussion: Study of geochemical element concentrations in 7 geological units showed that the Ca, Fe, Mg and Al elements have the highest and B and Co have the lowest concentrations within the soil samples. Overall evaluation of classification algorithms in training stage showed that the discriminant analysis, random forest, k nearest neighbor and support vector machines with linear, polynomial, multiple and RBF kernels with maximum values of the coefficient of determination (R^2=1) and minimum values of the mean squared error (RMSE=0) are the most accurate algorithms in sediment source separation but the regression trees method has the worst performance. Also, at testing stage, the support vector machines with RBF kernel was the most accurate and the classification trees with maximum error rate was the most inaccurate algorithm. Also, entrance of geochemical and granulometric variables lead to the highest and lowest accuracy in the sediment source separation, respectively. Using the geochemical variables for the separation of sediment sources, types of support vector machines, nearest neighbor analysis, discriminant analysis and the random forest algorithm had the highest coefficients of determination and lowest error values in the training and testing stages. By entering the lithological variables, the random forest algorithm had the highest accuracy for the sediment sources classification in the training and testing

1. Assistant Professor University of torbat heydarieh - - Razavi-khorasan, University of torbat heydarieh - me.bashiri@yahoo.com
2 M.Sc. Student University of torbat heydarieh - - Razavi-khorasan, University of torbat heydarieh
3 Assistant Professor Ferdowsi University of mashhad - - Razavi-khorasan, Ferdowsi University of Mashhad
DOI: 10.22052/deej.2018.7.19.49

stages and the discriminant analysis and support vector machines were located thereafter. Finally, fitting the classification algorithms using granulometric variables showed that the support vector machines had highest accuracy in the training and testing stages of models and the random forest and nearest neighbor analysis were ranked thereafter.

Conclusion: Totally, due to the proper accuracy and performance of data mining classifier algorithms, application of these methods in the natural sciences is suggested especially in the large amounts of data. These algorithms are used to find patterns in large sets of data and help classify new information. Especially, the support vector machines that are supervised classifier algorithms and besides that, in the natural sciences have successful results. In the watershed management considering the time and cost, sediment source ascriptions are difficult to obtain using monitoring techniques, but data mining procedures, have emerged as a potentially valuable alternative. Therefore, application and evaluation of these methods are suggested for further studies and natural sciences data.

Keywords: Classification algorithms, Element density, Nodeh watershed, Sediment source ascription.