محیط‌شناسی، دورهٔ ٤١، شمارهٔ ٤، زمستان ١٣٩٤، صفحهٔ ٩٤٢-٩٢٣

تعیین مناطق آسیب‌پذیر آبخوان دشت ملکان به نیترات با استفاده از روش جنگل تصادفی

حسین نوروزی[1]، اصغر اصغری مقدم[2]، عطاالله ندیری[3]*

۱. کارشناسی ارشد هیدروژئولوژی، دانشکدهٔ علوم طبیعی، دانشگاه تبریز Hosseinnorouzi168@yahoo.com
۲. استاد هیدروژئولوژی، دانشکدهٔ علوم طبیعی، دانشگاه تبریز Moghaddam@tabrizu.ac.ir
۳. استادیار هیدروژئولوژی، دانشکدهٔ علوم طبیعی، دانشگاه تبریز

تاریخ وصول مقاله: ١٣٩٤/١/٣ تاریخ پذیرش مقاله: ١٣٩٤/٧/١١

چکیده

به دلیل وجود آنومالی نیترات در آب زیرزمینی دشت ملکان، ۲۷ نمونه از منابع آب زیرزمینی در شهریور سال ۱۳۹۳ جمع آوری و در آزمایشگاه آبشناسی دانشگاه تبریز تجزیهٔ هیدروشیمیایی شد. در مطالعهٔ حاضر روش جنگل تصادفی (RF)، که روشی یادگیری مبتنی بر دسته‌ای از درخت‌های تصمیم است، برای ارزیابی آسیب‌پذیری پیشنهاد شده است. روش RF نسبت به روش‌های دیگر دارای مزایایی مانند دقت پیش‌بینی بالا، توانایی زیاد در تعیین متغیرهای مهم در پیش‌بینی و ماهیت غیرپارامتری است. در این مقاله عملکرد روش RF برای مدل‌سازی پیش‌بینی آسیب‌پذیری ویژهٔ آبخوان دشت ملکان با استفاده از چهار دسته از داده‌ها شامل مدل A با تمام متغیرها، مدل B با متغیرهای مربوط به خصوصیات آبخوان، مدل C با متغیرهای نیروهای محرک و مدل D با متغیرهای مربوط به روش دراستیک ارزیابی شد. مدل‌های A و B با کمترین MSE به ترتیب برابر ۰/۰۱۲ و ۰/۰۱۳ و بیشترین AUC به‌منزلهٔ روش‌های مناسب برای آسیب‌پذیری آب زیرزمینی به آلودگی نیترات انتخاب شدند و مدل‌های C و D با داشتن بیشترین MSE به ترتیب برابر با ۰/۰۱۵ و ۰/۰۲۶ و کمترین AUC به‌منزلهٔ روش‌های نامناسب شناخته شدند. مدل A که دقیق‌ترین مدل شناخته شد ۴۴ درصد از منطقه را در محدودهٔ آسیب‌پذیری زیاد شناسایی کرد.

کلیدواژه

آب زیرزمینی، آسیب‌پذیری، جنگل تصادفی، دشت ملکان، نیترات.

۱. سرآغاز

نبود شناخت صحیح و درک میزان آسیب‌پذیری آب‌های زیرزمینی ممکن است سبب بروز آلودگی‌های شدید در این منابع شود و چه بسا دیگر نمی‌توان از این منابع استفاده کرد و برای رفع آلودگی و مصرف مجدد باید وقت و هزینهٔ زیادی صرف شود. افزایش فعالیت‌های انسانی در سال‌های اخیر آلاینده‌های بسیاری را وارد محیط زیست کرده و آب زیرزمینی را به‌منزلهٔ منبعی طبیعی، در معرض آلودگی صنعتی و کشاورزی قرار داده است. مدیریت مناسب آب‌های زیرزمینی، به‌خصوص در مناطق خشکی مانند ایران، بسیار ضروری است و این نگرانی با گسترش سریع کشاورزی، صنعت، افزایش جمعیت و تغییرات

* نویسنده مسئول: Email: Nadiri@tabrizu.ac.ir

اقلیمی که بر کیفیت و کمیت منابع آب زیرزمینی اثر می‌گذارند بیشتر می‌شود. از این رو، آلودگی آب‌های زیرزمینی می‌تواند سلامت انسان را به خطر اندازد. متأسفانه مسائل مربوط به تنزل کیفی آب زیرزمینی در بیشتر موارد به‌سختی می‌تواند مشاهده شود، به علت اینکه وقتی می‌توان به عوامل آلودگی پی برد که آثار آلودگی در چاه پمپاژ مشخص شود. سرعت کم آب زیرزمینی و ناهمگنی موجود در آب زیرزمینی موجب می‌شود که آشکارشدن مواد آلاینده خیلی دیر صورت گیرد (Asghari moghaddam et al., 2010). با توجه به اینکه بیشتر آلودگی آب‌های زیرزمینی به دست عوامل انسانی صورت می‌گیرد، اقدامات زیست‌محیطی باید به طور عمده در پیشگیری از آلودگی متمرکز شود. یکی از راه‌های مناسب برای جلوگیری از آلودگی آب‌های زیرزمینی شناسایی مناطق آسیب‌پذیر آبخوان و مدیریت کاربری اراضی است (Zabet, 2002). آسیب‌پذیری آب زیرزمینی نوعی خصوصیت نسبی بدون بعد و غیر قابل اندازه‌گیری است و به ویژگی‌های آبخوان، محیط زمین‌شناسی و هیدروژئولوژی بستگی دارد (Antonakos and Lambrakis, 2007; Thapinta and hudak, 2003). دو نوع کلی از آسیب‌پذیری مطرح است: اولین مورد، آسیب‌پذیری ذاتی است که حساسیت آبخوان نیز نامیده می‌شود و با ویژگی‌های آبخوان، مواد پوشاننده و شرایط هیدروژئولوژیکی تعیین می‌شود. نوع دوم، آسیب‌پذیری خاص است که با ویژگی‌های ذاتی آبخوان همچنین، عوامل انسانی مانند کاربری اراضی و نوع آلاینده مشخص می‌شود (Fijani et al., 2013). ارزیابی آسیب‌پذیری آب زیرزمینی نیاز به روش‌هایی دارد که بر پایهٔ شناخت هیدروژئولوژیکی و کاربرد مدل‌های پیش‌بینی قرار دارند. تصمیم‌گیری در خصوص آسیب‌پذیری مناطق و تجزیه و تحلیل مؤثر داده‌های جمع‌آوری‌شده بدون مدلی مناسب و کارامد امکان پذیر نیست (Nadiri et al., 2013). تجزیه و تحلیل اطلاعات با هدف ایجاد مدل پیش‌بینی مکانی دقیق و قابل توضیح، برای ادغام و ترکیب متغیرهای مربوط به آسیب‌پذیری به آلودگی، دارای اهمیت ویژه‌ای است (Tilahum and Merkel, 2009). تاکنون روش‌هایی که برای ارزیابی آسیب‌پذیری ابداع شده‌اند بیشتر بر شواهد آلودگی بنا شده‌اند و به طور نسبی از داده‌های کمتری استفاده می‌کنند (Rodriguez, 2012). در سالیان اخیر، ابزارهای یادگیری[1] و روش‌های جدیدی برای حل برخی از مشکلات فوق‌الذکر ارائه شده‌اند و به طور گسترده‌ای استفاده می‌شوند که این روش‌های یادگیری جدید با بهره‌گیری از رگرسیون‌های گروهی در حال پیدایش‌اند. یکی از انواع روش‌های یادگیری که از الگوریتم‌های پایه برای پیش‌بینی چندگانهٔ تکراری[2] استفاده می‌کند جنگل تصادفی نامیده می‌شود (Breiman, 2001; Friedl et al., 1999). این روش در طبقه‌بندی پوشش زمین[3] که از داده‌های سنجش از دور حاصل می‌شود و نیز در مسائل الکترونیک و پزشکی کاربرد فراوان دارد (Pal, 2005; Sesnie et al., 2008; Ko et al., 2011). بولستیکس و همکاران در سال ۲۰۱۲ قابلیت‌های روش RF را در زمینهٔ آنالیز داده‌های بیواینفورماتیک بررسی کردند و بسیاری از جنبه‌های RF را مورد بحث قرار دادند. همچنین در این تحقیق پیاده‌سازی‌های مختلف RF معرفی و مقایسه شده است. گیسالسون در سال ۲۰۰۶ الگوریتم RF را به منظور طبقه‌بندی پوشش اراضی با استفاده از داده‌های سنجش از دور بررسی کرد. این تحقیق نشان داد که روش RF تا حد زیادی دقت طبقه‌بندی پوشش اراضی را بهبود می‌بخشد و احتمال برازش اضافی در آن وجود ندارد. همچنین روش RF پتانسیل لازم به‌منزلهٔ ابزار مدل مکانی برای ارزیابی آسیب‌پذیری در مباحث زیست‌محیطی و منابع آب را داراست (Booker and Snelder, 2012). در سال ۲۰۱۲، یو و همکاران عملکرد رگرسیون لجستیک، رگرسیون منطقی، درخت کلاس‌بندی رگرسیونی و RF را بررسی کردند. در این مطالعه مشخص شد که روش RF بهترین عملکرد را در میان روش‌های به‌کاررفته داشت. پهلوان و

همکاران در سال ۲۰۱۴ از روش RF برای به‌روزرسانی نقشه‌های خاک در شمال ایران استفاده کردند و نتیجهٔ پژوهش این محققان نشان داد که مجموعهٔ نقشه‌های به‌روزشدهٔ خاک با RF، حدود ۱۳/۴ درصد دقیق‌تر از مجموعهٔ نقشه‌های معمولی خاک است. RF روشی گروهی است که چند الگوریتم درختی را برای تولید پیش‌بینی‌ای مکرر از هر پدیده ترکیب می‌کند. RF می‌تواند الگوهای پیچیده را یاد بگیرد و ارتباط غیرخطی بین متغیرهای توضیحی و متغیرهای وابسته را در نظر بگیرد. همچنین می‌تواند انواع مختلف داده‌ها را در تجزیه و تحلیل بگنجاند و ترکیب کند که این هم به علت نبود توزیع پیش‌فرض‌ها (توزیع نرمال) دربارهٔ داده‌های استفاده‌شده است. RF هزاران متغیر ورودی را بدون حذف یکی از آن‌ها پذیراست و اجرا می‌کند؛ همچنین می‌تواند برآوردی را از اینکه کدام متغیر در پیش‌بینی مدل مهم است ارائه دهد (Rodriguez, 2012). RF نسبت به شبکه‌های عصبی مصنوعی در برابر گیرافتادن در مینیمم محلی و داده‌های پرت حساسیت کمتری دارد و می‌تواند تخمین بهتری از پارامترها را داشته باشد. RF اهمیت نسبی متغیرها را ارزیابی می‌کند همچنین قادر به انتخاب متغیرهای مهم است و در عین حال پارامترسازی آن نسبت به روش‌های دیگر مانند شبکه‌های عصبی محاسبات ساده‌تری دارد (Rodriguez, 2012). در مطالعهٔ حاضر، مدل RF برای آبخوان دشت ملکان با کمک پایگاهی از داده‌های سیستم اطلاعات جغرافیایی شامل داده‌های خصوصیات هیدروژئولوژیکی آبخوان، فاصله از فعالیت‌های انسانی و منابع بالقوهٔ تغذیه، متغیرهای سنجش از دور و خصوصیات فیزیکی و شیمیایی اندازه‌گیری‌شده در صحرا، توسعه داده شده است و پتانسیل RF برای ایجاد نقشهٔ آسیب‌پذیری با توجه به معیارهای مختلف، که وابسته به تغییرات پارامترهای الگوریتم و دقت نقشه‌هاست، بررسی می‌شود. این روش تاکنون در مباحث زمین‌شناسی و هیدروژئولوژیکی به صورت کامل ارزیابی نشده است.

هدف کلی از این تحقیق توسعهٔ مدل پیش‌بینی دقیقی برای آسیب‌پذیری آب زیرزمینی نسبت به آلودگی نیترات با استفاده از داده‌های مکانی مربوط به خصوصیات ذاتی و خاص آبخوان، نیروهای محرک و خصوصیات فیزیکی ـ شیمیایی است. مجموعه‌ای از اهداف جزئی مطالعهٔ حاضر عبارت‌اند از: الف) ارزیابی عملکرد مدل RF برای تخمین آسیب‌پذیری آب زیرزمینی با استفاده از داده‌های ذکرشده در بالا، ب) ساختن مدل‌های مختلف بر پایهٔ ارزیابی آسیب‌پذیری ویژه برای بررسی بهترین مدل پیش‌بینی، ج) استفاده از روش RF برای تعیین اهمیت پیش‌بینی‌کننده‌های آلودگی آب زیرزمینی به نیترات، د) استفاده از انتخاب ویژگی RF برای کاهش ابعاد، بهتر قابل توضیح‌بودن و افزایش دقت، هـ) پیش‌بینی مکانی غلظت نیترات بیش از ۵۰ میلی‌گرم در لیتر (حد مجاز) در آب‌های زیرزمینی با توجه به نقشه‌های طبقه‌بندی.

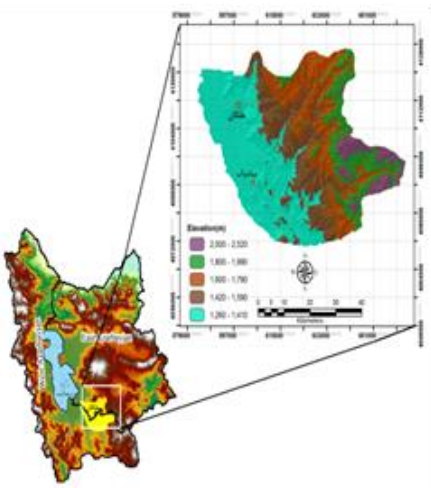
۲. مواد و روش‌ها

۱.۲. منطقهٔ مورد مطالعه

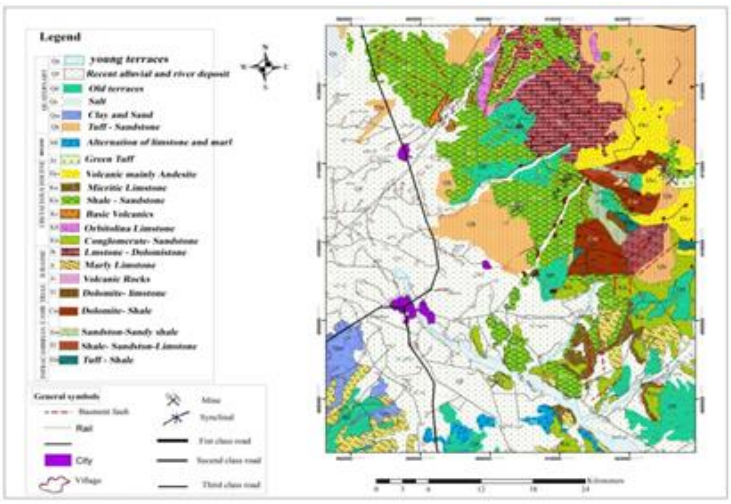
دشت ملکان با وسعتی تقریباً برابر با ۴۵۰ کیلومتر مربع در جنوب استان آذربایجان شرقی و در جنوب‌شرق دریاچهٔ ارومیه واقع شده است (شکل ۱) و جزء زون زمین‌ساختاری البرز ـ آذربایجان محسوب می‌شود. از نظر توپوگرافی دشت ملکان در محدودهٔ ارتفاعی ۱۳۲۰ متر قرار دارد. در واقع اختلاف ارتفاع از رأس دشت تا انتهای دشت ۴۵ متر است. دشت ملکان از نظر تقسیمات طبیعی در حوضهٔ آبریز دریاچهٔ ارومیه قرار دارد و این حوضه بر اساس روش تجربی آمبرژه (Emberger, 1952) و با استفاده از آمار ایستگاه ملکان دارای اقلیم سرد و نیمه‌خشک است. منطقهٔ ملکان دارای سازندهای زمین‌شناسی مختلفی است. سازند لالون که در قسمت شرق و سازند روته در دورهٔ پرمین در جنوب منطقه برونزد دارد، سازند شمشک در قسمت شرق و شمال‌شرقی مشهود است و سازند لار مربوط به دورهٔ تریاس و ژوراسیک در

قسمت شمال و شمال‌شرق منطقه دیده می‌شود. همان طور که در شکل ۲ نشان داده شده است، بخش اعظمی از مساحت منطقه در بخش غربی مربوط به رسوبات آبرفتی دورهٔ کواترنر است و بخش کمی از آن‌ها در بخش شمال‌غرب پهنه‌های لسی ـ نمکی دارند. در قسمت جنوب و جنوب‌شرقی دشت ملکان مجموعهٔ رسوبات آهک و مارن ـ ژیپس دیده می‌شود که مربوط به دورهٔ کرتاسه و مجموعه سنگ‌های پیروکلاستیک و رس‌سنگ‌ها مربوط به دورهٔ پلیوسن‌اند. در قسمت شمال و شمال‌غربی دشت ملکان، سنگ‌های آهکی تا تودهٔ خاکستری روشن مربوط به دورهٔ تریاس و ژوراسیک و تودهٔ آهک‌های کرتاسه به چشم می‌خورد. آبخوان دشت از نوع آزاد است که اکثراً از پادگانه‌های آبرفتی قدیمی، پادگانه‌های آبرفتی جدید، مخروط‌افکنه‌ها و رسوبات رودخانه‌ای تشکیل یافته است و مواد اصلی تشکیل‌دهندهٔ آبخوان رسوبات ماسه، سیلت و رس است. مطالعات نشان می‌دهد که سطوح ایستابی دشت در ایستگاه‌های مورد مطالعه، از سمت واحد مخروط‌افکنه تا واحد شوره‌زار ساحلی افزایش می‌یابد؛ بدین معنی که از رأس مخروط‌افکنه به سمت پای دشت عمق سطح ایستابی آب‌های زیرزمینی کاهش می‌یابد. منطقهٔ ملکان یکی از قطب‌های تولید انگور کشور است که برای افزایش بازدهی آن از مقادیر زیادی کودهای شیمیایی و حیوانی استفاده می‌شود؛ همچنین به دلیل وجود چاه‌های جذبی و نبود شبکهٔ فاضلاب در مناطق شهری و روستایی، میزان نیترات در بسیاری از منابع آب زیرزمینی دشت بالاتر از حد استاندارد جهانی است. شکل ۳ میزان نیترات اندازه‌گیری‌شده در سال ۱۳۹۳ را نشان می‌دهد.

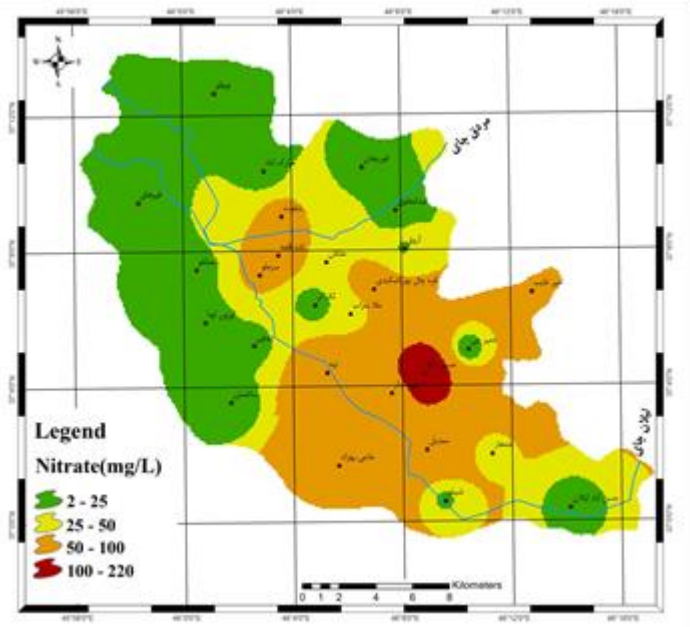
بر اساس نقشه‌های هم‌ضخامت رسوبات آبرفتی، لاگ‌های حفاری و داده‌های ژئوفیزیکی، در قسمت‌های بالایی دشت ذرات تشکیل‌دهندهٔ سفره دانه‌درشت‌اند و هر چه به سمت مرکز دشت و نواحی خروجی و به سمت دریاچهٔ ارومیه نزدیک می‌شویم رسوبات در این مناطق دانه‌ریزتر می‌شوند. عمق کم سطح آب زیرزمینی در قسمت‌های کم‌ارتفاع دشت باعث تبخیر زیاد و افزایش شوری بیش از اندازهٔ آب‌های زیرزمینی می‌شود، ضمن اینکه ریزدانه‌بودن رسوبات نیز باعث می‌شود که آب در اثر نیروی کاپیلاری بالا آید و به تبع آن میزان تبخیر نیز افزایش یابد (Lehmann and Or, 2009).



شکل ۱. موقعیت منطقهٔ مورد مطالعه



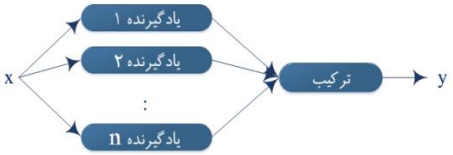
شکل ۲. نقشهٔ زمین‌شناسی منطقهٔ مورد مطالعه



شکل ۳. نقشهٔ نیترات آب زیرزمینی دشت ملکان (سال ۱۳۹۳)

۲.۲. روش‌شناسی

یکی از ابزارهای کارامد در مسائل مربوط به تخمین متغیرهای هدف و یا طبقه‌بندی الگوها درخت تصمیم است. درخت تصمیم فضای ورودی را به مجموعه‌هایی از نواحی مجزا تقسیم می‌کند و پاسخی را به هر ناحیه اختصاص می‌دهد. در حالت ساده، این پاسخ در مسائل رگرسیونی می‌تواند بر اساس میانگین مقادیر هدف، مرتبط با الگوهای قرارگرفته در هر ناحیه تعیین شود. پاسخ اختصاص‌یافته به هر ناحیه بر اساس میانگین مقادیر هدف با الگوهای یادگیری قرارگرفته در هر ناحیه متناظر می‌شود. به طور کلی، درخت تصمیم منفرد مستعد برازش اضافی[۴] است و قدرت تعمیم‌پذیری کمی دارد. در هنگام تشکیل درخت تصمیم، تغییر کوچکی در الگوهای یادگیری می‌تواند باعث تغییرات اساسی در ساختار آن درخت شود (Quinlan, 1986). ترکیب درخت‌های تصمیم را، که از این مشکل جلوگیری می‌کند، روش‌های دسته‌جمعی می‌گویند. شکل ۴ الگوی سادهٔ روش دسته‌جمعی معمولی را نشان می‌دهد. مجموعه‌ای دسته‌جمعی شامل تعدادی یادگیرنده است که به آن‌ها یادگیرنده‌های پایه گفته می‌شود. یادگیرنده‌های پایه معمولاً به وسیلهٔ الگوریتم یادگیری پایه، که می‌تواند درخت تصمیم، شبکهٔ عصبی یا الگوریتم‌های یادگیری دیگر باشد، از دادهٔ آموزشی ساخته می‌شوند. قابلیت تعمیم یک مجموعه اغلب قوی‌تر از یادگیرنده‌های پایه است. در واقع روش‌های مجموعه‌ای بیشتر به دلیل توانایی در تقویت یادگیرنده‌های ضعیف مقبول‌اند (Schapire, 1990). از این رو به یادگیرنده‌های پایه یادگیرنده‌های ضعیف نیز گفته می‌شود. برای فائق‌آمدن بر مشکلات یادگیرنده‌های پایه، الگوریتم جنگل تصادفی که روشی یادگیری مبتنی بر دسته‌ای از درخت‌های تصمیم است پیشنهاد شده است.



شکل ۴. یک روش دسته‌جمعی معمولی

۳.۲. رگرسیون و الگوریتم جنگل تصادفی

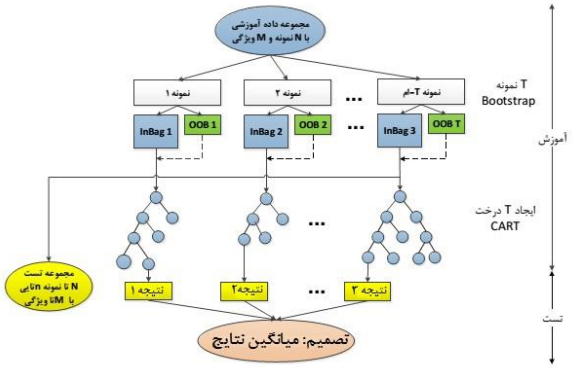
در الگوریتم RF، برای تشکیل هر درخت، دستهٔ متفاوتی از الگوهای موجود، با درنظرگرفتن جایگزینی دوبارهٔ هر

الگوی انتخاب‌شده، انتخاب می‌شوند. اندازهٔ این دستهٔ نمونه‌برداری‌شده برابر تعداد کل الگوهای موجود خواهد بود. RF در سال ۲۰۰۱ به دست Breiman به‌منزلهٔ روشی از توسعهٔ جدید درخت‌های تصمیم‌گیری بیان شد که پیش‌بینی چندین الگوریتم منفرد را با هم با استفاده از قوانین مبتنی ترکیب می‌کند. اصول کلی تکنیک‌های آموزش گروهی بر پایهٔ این فرض است که دقت آن‌ها از دیگر الگوریتم‌های آموزشی بالاتر است، چون ترکیبی از چند مدل پیش‌بینی دقیق‌تر از یک مدل است و گروه‌ها قدرت مجموعه‌های منفرد و منحصر به فرد از طبقه‌ها را بیشتر می‌کنند در حالی که، نقاط ضعف طبقه‌ها را در همان زمان کاهش می‌دهند (Kotsiantis and Pintelas, 2004). هدف از این تحقیق بررسی آسیب‌پذیری آبخوان دشت ملکان به آلودگی نیترات است که فقط در حالت رگرسیونی بحث می‌شود. درخت رگرسیونی (RT) مجموعه‌ای از شرایط یا محدودیت‌ها را بیان می‌کند که به صورت سلسله‌مراتبی سازمان یافته‌اند و به حالت متوالی از گره ریشه[۵] به سمت پایین رشد می‌کنند و به گره‌های پایانی[۶] یا گره‌های برگ[۷] می‌رسند (Breiman, 1984; Quinlan, 1993). به منظور به‌وجودآوردن درخت رگرسیونی از پارتیشن‌بندی بازگشتی و رگرسیون‌های چندگانه استفاد می‌شود. از گره ریشه، فرایند تصمیم در هر گره داخلی، طبق قانون درختی تا زمانی که شرط توقف قبلی تعیین‌شده به دست آید، تکرار می‌شود. هر یک از گره‌های نهایی یا برگ‌ها به مدل رگرسیونی ساده، که فقط در گره به کار برده می‌شوند، متصل می‌شوند. زمانی که فرایند فراخوانی درخت به پایان برسد هرس‌کردن یا پرونینگ[۸] با هدف بهبود ظرفیت تعمیم درخت‌ها به وسیلهٔ کاهش پیچیدگی ساختار به کار برده می‌شود. تعداد نمونه‌ها در گره‌ها می‌تواند به‌منزلهٔ معیار پرونینگ در نظر گرفته شود. برای جلوگیری از تطابق بیش RTsهای مختلف، RF تنوع درختان را از طریق درست‌کردن زیرمجموعه‌های[۹] مختلف از داده‌های آموزشی کم می‌کند که اصطلاحاً کیسه‌بندی[۱۰] نامیده می‌شود. اصطلاح Bagging از مخفف **B**ootstrap **Agg**regat**ing** به دست آمده است (Breiman, 1996). Bagging روشی است که از طریق نمونه‌برداری دوبارهٔ تصادفی از مجموعه داده‌های اصلی و همراه با جایگزینی، برای ایجاد داده‌های آموزشی، به کار می‌رود و در این مرحله هیچ‌کدام از داده‌های انتخاب‌شده از نمونه‌های ورودی را برای تولید زیرمجموعهٔ بعدی حذف نمی‌کند، بدین ترتیب واریانس نیز کاهش می‌یابد. از این رو برخی از داده‌ها ممکن است بیش از یک‌بار در شاخه‌های آموزشی استفاده شوند در حالی که، برخی از داده‌های دیگر که در مدل‌سازی مؤثر نیستند هرگز استفاده نمی‌شوند. بنابراین ثبات بیشتری برای مدل به دست می‌آید و مدل را در برابر تغییرات جزئی در داده‌های ورودی قابل اعتمادتر می‌کند و دقت پیش‌بینی آن را افزایش می‌دهد (Breiman, 2001). از سوی دیگر، هنگامی که RF یک درخت رشد ایجاد می‌کند از بهترین متغیرها یا نقاط تقسیم در داخل زیرمجموعه‌های متغیرها استفاده می‌کند که به صورت تصادفی از مجموعه‌های کلی متغیرهای ورودی انتخاب می‌کند بنابراین قدرت هر RT منفرد را کاهش می‌دهد و میزان تطابق را پایین می‌آورد و بدین صورت خطای کلی مدل را کاهش می‌دهد (Breiman, 2001). این روش یک متا الگوریتم است که برای بهبود یادگیری ماشین رده‌بندی و مدل‌های پسرفتی بر حسب پایداری و دقت رده‌بندی است. این روش همچنین واریانس را کاهش می‌دهد و به دوری از اورفیتینگ کمک می‌کند. اگرچه این روش در درخت تصمیم به کار می‌رود، اما می‌تواند در هر نوع مدل استفاده شود. Bagging حالتی مخصوص از روند مدل میانگین است. یکی دیگر از ویژگی‌های خوب RF این است که درختان RF بدون پرونینگ یا هرس‌کردن رشد می‌کنند و در این روش آموزش بیش از اندازه بر دقت مدل تأثیری نمی‌گذارد که آن را از دیدگاه محاسباتی سبک‌تر می‌کند. علاوه بر این، آن دسته از نمونه‌هایی که در آموزش درختان در فرایند Bagging انتخاب نمی‌شوند شامل بخشی از

زیرمجموعه‌هایی می‌شوند که الگوهای خارج از کیسه[11](oob) نامیده می‌شوند. و این قسمت در روش RF می‌تواند برای ارزیابی عملکرد مدل استفاده شود (Peters et al., 2007). به این ترتیب، RF می‌تواند تخمین غیرمرتبط داخلی از خطای تعمیم را محاسبه کند بدون اینکه از زیرمجموعه‌های داده‌های خارجی استفاده کند (Breiman, 2001). روند کلی الگوریتم RF به صورت ساده در شکل ۵ نشان داده شده است. در این روش بردار تصادفی Θ_k که مستقل از بردارهای تصادفی $\Theta_1 \ldots. \Theta_{k-1}$ است، برای درخت Kth تولید می‌شود. همچنین، همهٔ بردارها توزیع یکسانی دارند. درخت رگرسیونی با استفاده از مجموعه داده‌های آموزش و Θ_k رشد می‌کند و نتیجهٔ مجموعه درخت‌های Kبرابر، $\{h_1(x), h_2(x) \ldots h_k(x)\}$ است که در این جا $h_k(x) = h(x, \theta_k)\ , x = \{x_1, x_2, \ldots x_p\}$ است، همچنین این بردارها بردار ورودی P بعدی‌اند که یک جنگل را تشکیل می‌دهند. خروجی‌های K تولیدشدهٔ گروهی، مربوط به هر درخت برابر $\hat{y}_1 = h_1(x), \hat{y}_2 = h_2(x), \ldots \hat{y}_k = h_k(x)$ که $\hat{y}_k$ خروجی درخت Kth است. برای به‌دست‌آوردن خروجی نهایی، متوسط همهٔ پیش‌بینی‌های درخت‌ها محاسبه می‌شود. خطای پیش‌بینی نیز بر اساس نمونه‌های Out Of Bag طبق فرمول زیر محاسبه می‌شود.

$$MSE \simeq MSE^{OOB} = n^{-1}\sum_{i=1}^{n}\left[\hat{y}(x_i) - y_i\right]^2 \qquad (۱)$$

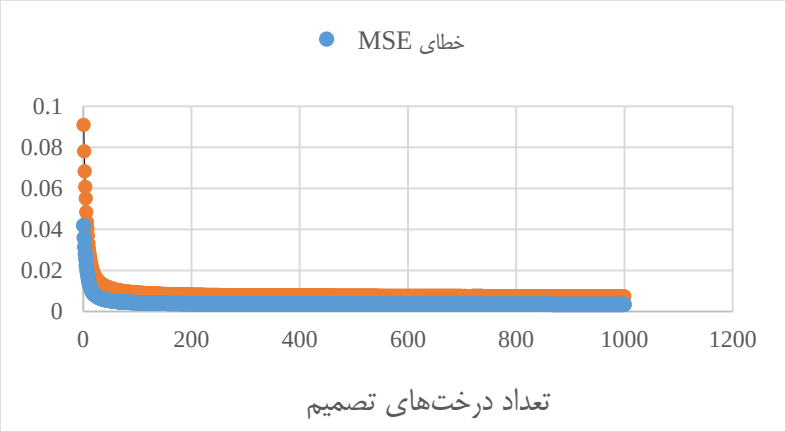
اگر بخواهیم روش RF برای طبقه‌بندی را به طور خلاصه بیان کنیم، بدین صورت است که: در ابتدا T نمونه خودراه‌انداز از دادهٔ آموزشی بیرون کشیده می‌شود سپس، از هر نمونه خودراه‌انداز β یک درخت طبقه‌بندی و رگرسیون (CART) هرس‌نشده ایجاد می‌شود که برای انشعاب در هر گره CART، تنها یکی از M ویژگی انتخاب‌شده به صورت تصادفی استفاده می‌شود. در نهایت، خروجی طبقه‌بندی بر اساس یک نتیجه میانگین از پیش‌بینی‌های تمام درخت‌های منفرد آموزش‌دیده به دست می‌آید. یک مجموعه دادهٔ خودراه‌انداز مجموعه‌ای از نقاط انتخابی به طور تصادفی است که با جاگذاری از مجموعه دادهٔ آموزشی بیرون کشیده شده است (Duda et al., 2011). برای اینکه همیشه اندازهٔ نمونهٔ آموزشی اولیه ثابت بماند، مجموعه دادهٔ خودراه‌انداز نسخه‌ای کپی از نقاط را به کار می‌برد. شایان ذکر است که الگوریتم RF به دلیل ارزیابی درونی نتایج هر طبقه‌بندی درختی که در داخل خود دارد و با وزن‌دهی به نتایج هر درخت می‌تواند نتایج صحیحی را تولید کند.



شکل ۵. فلوچات RF برای رگرسیون
برگرفته از: Guo et al., 2011

گرچه اطلاعات زیاد برای مدل‌سازی ممکن است مفید باشند، اما افزایش تعداد پارامترهای ورودی پیچیدگی‌های اضافی و افزایش زمان محاسبات و مشکلات ابعادی را به سیستم تحمیل می‌کند (Bellman, 2003). در تلاش برای رسیدن به تقسیمی بهینه، تصمیم‌ساز[12] می‌تواند به وسیلهٔ مقدار بسیار عظیمی از داده‌های مختلف اطلاعات و شرایط موجود در منطقه را به‌سرعت تحت‌الشعاع قرار دهد. تعداد زیادی از متغیرهای وابسته به خواص و رفتار سیستم آب زیرزمینی و نیروهای راندگی می‌توانند توانایی مدل را افزایش دهند. این ابعاد بالا در مجموعهٔ داده‌ها می‌تواند باعث کاهش دقت مدل شود. برای جلوگیری از این خطاها و ابعاد بالای داده‌ها انتخاب ویژگی[13] (FS) به کار برده

می‌شود. FS روشی برای انتخاب زیرمجموعه‌های پارامترهای مربوطه برای آموزش بهتر مدل است (Guyon and Elisseeff, 2003). در مطالعات آب زیرزمینی، از تعداد زیادی متغیرهای وابسته به خصوصیات فیزیکی و شیمیایی آبخوان استفاده می‌شود که برخی از آن‌ها می‌توانند مرتبط یا برخی غیرمرتبط باشند (Dixon, 2009). بنابراین در ارزیابی آسیب‌پذیری ویژهٔ آبخوان، انتخاب متغیرهای توضیحی که ارتباط بیشتری با آلاینده دارند خیلی مهم است. FS با بالابردن سرعت فرایند آموزش، افزایش قابلیت تعمیم، کاهش اثر ازبین‌رفتن ابعاد و افزایش قابلیت تفسیر دقت مدل‌های پیش‌بینی را افزایش می‌دهد. روش‌های زیادی برای FS بیان شده است. روش معمول FS روش آماری چندمتغیره است که به سبب کاهش ابعاد داده‌ها، مؤلفه‌های اصلی جایگزین پارامترهای اولیه می‌شوند (Critto et al., 2003) که در مطالعهٔ حاضر روش RF که بر اساس درخت‌های گروهی است ترجیح داده شده است. در این روش همان طور که در شکل ۶ نشان داده شده است، با افزایش درختان، خطای تعمیم همیشه همگرا می‌شود، از این رو آموزش زیاد در این روش مشکلی ندارد و تعداد درختان زمانی که خطا همگرا شده است می‌تواند ثابت باشد. با این حال انتخاب متغیر m بسیار مهم است. در هر گره، زیرمجموعهٔ متغیرهای پیش‌بینی‌کنندهٔ m از یک تا ماکزیمم مشخص می‌شوند. با وجود رشد درخت‌ها و انتخاب تصادفی متغیرها، مقدار باقیماندهٔ m می‌تواند ثابت باقی بماند. بنابراین، تعریف یا انتخاب پارامتر m تطابق و قدرت هر درخت منفرد را تحت تأثیر قرار می‌دهد و در نتیجه بر خطای تعمیم و دقت مدل پیش‌بینی مؤثر است (Breiman, 2001).



شکل ۶. تأثیر تعداد درخت‌ها در میزان خطای MSE

۴.۲. طراحی اطلاعات و فراخوانی مدل آلایندهٔ نیترات

در مطالعهٔ حاضر از ۲۷ چاه مورد نظر با پراکندگی مناسب به طوری که بتوانند نمایانگر تغییرات هیدروشیمیایی کل دشت باشند در شهریور سال ۱۳۹۳ نمونه‌برداری انجام شد و بعضی پارامترهای فیزیکوشیمیایی (pH، EC، T) در محل نمونه‌برداری اندازه‌گیری شدند. آنالیزهای هیدروشیمیایی در آزمایشگاه آب‌شناسی دانشگاه تبریز انجام شد و مقادیر نیترات نمونه‌های برداشت‌شده به دست آمدند. جدول ۱ مقادیر نیترات اندازه‌گیری‌شده را نشان می‌دهد. همچنین متغیرهای دیگری مانند نیروهای راندگی[۱۴] (بعضی از فعالیت‌هایی که به صورت نقطه‌ای یا غیرنقطه‌ای باعث تولید خطر می‌شوند)، محیط خاک، متغیرهای سنجش از دور و پوشش زمین، شاخص تفاضلی نرمال‌شدهٔ پوشش گیاهی[۱۵] (NDVI) و خصوصیات ذاتی آبخوان به دست آورده شدند. جدول ۲ متغیرهای توضیحی

و متغیر پاسخ را که همان غلظت نیترات است بیان می‌کند که منابع داده‌ها و روش‌های برآورد نیز آورده شده‌اند. به منظور به‌دست‌آوردن متغیرهای پیوسته و استانداردشده، برای همهٔ منطقه مورد مطالعه، داده‌ها به فرمت رستری تبدیل شدند که در این مرحله از سه روش استفاده شد: الف) روش‌های زمین آماری برای به‌دست‌آوردن نقشه‌های هدایت الکتریکی، گرادیان هیدرولیکی، بافت خاک و ...، ب) محاسبات فاصلهٔ اقلیدسی رستری برای به‌دست‌آوردن لایه‌های پتانسیل منابع نقطه‌ای آلودگی، ج) طبقه‌بندی پوشش زمین از سنجش از دور و NDVI با تصاویر ماهواره‌ای. برخی از ویژگی‌های مهم خصوصیات ذاتی منطقه که بیان‌کنندهٔ انتخاب مدل رگرسیونی RF است به صورت زیر توضیح داده شده‌اند: ۱) شیب سطح زمین که نشان‌دهندهٔ این است که آیا رواناب روی سطح باقی خواهد ماند تا به نفوذ آلاینده به داخل زون اشباع کمک کند؛ ۲) سطح آب زیرزمینی و ضخامت زون وادوز نشان می‌دهند که آیا شست‌وشوی نیترات به داخل زون وادوز سریع اتفاق می‌افتد یا اینکه دیرتر صورت می‌گیرد (سطح آب زیرزمینی بالا باعث تسریع انتقال آلاینده به داخل آبخوان می‌شود)؛ ۳) بافت خاک به دلیل وجود رس یا شن و ماسه که انتقال نیترات و میزان نیتروژن ازدست‌رفته را کنترل می‌کند؛ ۴) میزان قابلیت انتقال آبخوان و هدایت هیدرولیکی آن که در سرعت آب و انتقال آلاینده نقش دارند؛ ۵) گرادیان هیدرولیکی که در انتقال آلاینده و میزان مهاجرت آلاینده نقش دارد؛ ۶) متغیرهای سنجش از دور نیز می‌توانند برای ارزیابی پتانسیل رواناب (نوع و میزان پوشش گیاهی موجود در منطقه بر میزان رواناب تأثیر دارد)، منابع پراکندهٔ آلاینده‌های کشاورزی و پتانسیل حذف نیترات کمک کنند.

جدول ۱. مقادیر نیترات در نمونه‌های آب زیرزمینی دشت ملکان

نیترات (میلی‌گرم بر لیتر)	محل نمونه‌برداری	شمارهٔ نمونه	نیترات (میلی‌گرم بر لیتر)	محل نمونه‌برداری	شمارهٔ نمونه
۹/۹۸	قوریجان	M15	۷۶/۲۶	امیرغایب	M1
۱۲/۳۸	میدانجلوق	M16	۸/۸۶	چپقلو	M2
۴۵/۱۰	ملکان	M17	۴/۷۷	مبارک آباد	M3
۱۴۵/۳۶	آروق	M18	۱۲/۹۵	دوچی	M4
۶۵/۱۲	بایقوت	M19	۲/۳۸	ساتلمش	M5
۶۰/۴۷	سرملو	M20	۷۳/۱۶	قره چال	M6
۲۰/۱۴	شبیلو	M21	۸۸/۲۵	تازه قلعه	M7
۱/۲۳	لک لر	M22	۵/۶۱	قیپچاق	M8
۲۸/۱۸	ملاسراب	M23	۲۲۲	حسین آباد	M9
۸۹/۹۲	مهماندار	M24	۲/۶۴	اوزون اوبا	M10
۴/۷۷	دمیرچی	M25	۱۰/۴۱	شعبانلو	M11
۹۶/۷۲	تپه اسماعیل آباد	M26	۷/۷۳	حسن آباد لیلان	M12
۱۲۵/۷	حاجی بهزاد	M27	۳۶/۲۱	قندهار	M13
			۱۱۱/۸۷	ممدیل	M14

جدول ۲. متغیرهای توضیحی و متغیر پاسخ، همراه با منبع و روش مورد استفاده در برآورد لایهٔ رستری

نوع متغیر	نام متغیر	داده‌های مورد استفاده	روش مورد استفاده
متغیرهای توضیحی خصوصیات سفره	تراز آب زیرزمینی ارتفاع سنگ کف آبخوان ضخامت اشباع قابلیت انتقال آبخوان و هدایت هیدرولیکی گرادیان هیدرولیکی ضخامت زون وادوز محیط خاک، خاک سطحی و محیط آبخوان لایهٔ تغذیه	اطلاعات ۳۰ پیزومتر موجود در منطقه اطلاعات ژئوفیزیک و گمانه‌های حفاری تراز سطح آب و ارتفاع سنگ کف استفاده از اطلاعات اطلاعات ۲۷ چاه پمپاژ انجام‌شده در منطقه منحنی‌های تراز آب زیرزمینی و فاصلهٔ بین منحنی‌ها عمق سطح آب زیرزمینی لاگ‌های حفاری منطقه لایه‌های شیب، بارندگی و نفوذپذیری خاک	روش‌های زمین آماری معکوس وزنی فاصله و کریجینگ پهنه‌بندی لایه‌ها با روش‌های زمین آماری و استفاده از روش Piscopo
متغیرهای توضیحی نیروهای محرک	فاصله از رودخانه‌ها، شهرها، باغات انگور، شهرک‌های صنعتی، محل‌های دفن زباله و کانال‌های آبیاری	تصاویر ماهواره‌ای و گوگل‌ارث، نقشه‌های با مقیاس کم و مختصات برداشتی نقاط انتخابی	اندازه‌گیری فاصلهٔ اقلیدسی با استفاده از GIS
متغیرهای توضیحی سنجش از دور	پوشش زمین شاخص NDVI شیب سطح زمین	تصاویر ماهواره‌ای سال ۲۰۱۴ ماه نوامبر Lndsat8	طبقه‌بندی پوشش زمین با داده‌های Landsat8
متغیر پاسخ	نیترات	آنالیز ۲۷ نمونهٔ برداشتی از آب زیرزمینی	پهنه‌بندی با استفاده از روش زمین‌آماری کریجینگ

۵.۲. طراحی مدل

برای مطالعهٔ گروه‌های مختلف متغیرها در پیش‌بینی پارامترهای زیرمجموعه‌های زیر استفاده شده‌اند: A) همهٔ متغیرها، B) متغیرهای مربوط به آسیب‌پذیری ذاتی و پارامترهای اندازه‌گیری‌شدهٔ مربوط به کیفیت آب زیرزمینی، C) نیروهای محرک D) متغیرهای مربوط به روش دراستیک. اولین مرحله در ایجاد مدل پیش‌بینی آلودگی نیترات تنظیم پارامترهای مدل است. به منظور تنظیم مقدار k، طوری که مقدار خطا همگرا شود و تخمین قابل اعتمادتر باشد، مدل از ۱ تا ۱۰۰۰ درخت ساخته شد. به دلیل اینکه با افزایش درخت‌ها میزان خطا کاهش می‌یابد، بنابراین تعداد ۱۰۰۰ درخت برای فراخوانی مدل استفاده شد. پارامتر m نیز به وسیلهٔ تغییر تعدادی متغیرهای تقسیم بین یک و ماکزیمم متغیرهای هر زیرمجموعه بهینه شد.

نتایج مدل به وسیلهٔ برآورد خطای oob ارزیابی شد. علاوه بر این، برای کاهش ابعاد و افزایش دقت و قابل تفسیربودن مدل روش FS به کار برده شد و متغیرهای مهم در پیش‌بینی نیز شناسایی شدند. برای فراخوانی مدل آلایندهٔ نیترات، داده‌های حاصل از لایه‌های رستری ۲۳ متغیر توضیح داده شد، با هم و با متغیر هدف استفاده شد. حد مجاز نیترات برای تمایز نقاط آلوده و غیرآلوده بر اساس اعلام سازمان بهداشت جهانی[۱۶] (WHO, 2009)، ۵۰ میلی‌گرم بر لیتر در نظر گرفته شد. غلظت نیترات نسبت به متغیر پاسخ جدید برای هر نمونهٔ تجربی دوباره مقیاس‌بندی شد. نمونه‌های با غلظت نیترات مساوی یا بیشتر از حد آستانه (۵۰ میلی‌گرم بر لیتر) مساوی یک و نمونه‌های کمتر از حد آستانه مساوی صفر در نظر گرفته شد. متغیرهای توضیحی (پیش‌بینی‌کننده) و متغیر پاسخ با هم در مجموعه‌ای از بردارهای ویژگی ورودی ترکیب شدند. این بردارها ورودی الگوریتم RF را تشکیل دادند و به‌عنوان بردارهای Input-feature شناخته شدند. متغیر پاسخ دوتایی (آلودگی نیترات) به‌منزلهٔ مقادیر هدف برای آموزش الگوریتم استفاده شدند.

۶.۲. ارزیابی دقت مدل‌سازی آلودگی نیترات

از آن‌جا که طبقه‌بندی RF از میانگین‌گیری و شمارش نتایج طبقه‌بندی‌کننده‌های پایهٔ ساده و متنوع نتیجهٔ طبقه‌بندی را تعیین می‌کند، قادر است با داده‌های دارای نویز نیز به طور صحیح‌تری در مقایسه با الگوریتم‌های دیگر مثل SVM و شبکه‌های عصبی، که در آن‌ها احتمال برازش اضافی در هنگام آموزش با داده‌های پرت بالاست، آموزش ببیند (Chehata et al., 2009). پیش‌بینی‌ها از ترکیب تعداد زیادی از پارامترهای ممکن ناشی می‌شود که به وسیلهٔ ریشهٔ میانگین مربعات خطا[۱۷] (RMSE) ارزیابی می‌شوند. مدل‌های بهینه برای همهٔ زیرمجموعه‌های RF مقایسه می‌شوند و مدلی که کمترین RMSE را داشته باشد برای پیش‌بینی انتخاب می‌شود که نرخ مثبت واقعی[۱۸] (TPR) با توجه به غلظت نمونه‌های برداشته‌شده از آب زیرزمینی و مقایسهٔ آن‌ها با غلظت پیش‌بینی‌شده سنجیده می‌شود. منحنی‌های عملکرد (کارایی) مدل با پلات‌کردن درصد مناطق آلوده در مقابل میزان نرخ موفقیت به دست می‌آیند. روش دیگر تجزیه و تحلیل مدل بر پایهٔ منحنی‌های مشخصهٔ عملکرد سیستم[۱۹] (ROC) است. منحنی‌های ROC به نحوی مشابه نرخ موفقیت‌اند که می‌توانند به وسیلهٔ نرخ مثبت واقعی کنترل شوند. منحنی‌های ROC با تغییر مقدار مجاز آستانه در برابر خروجی پیش‌بینی رسم می‌شوند. به طور کلی نتایج[۲۰] FPR در محور X در مقابل TPR در محور Yرسم می‌شود. هر یک از نتایج آستانه در جفت FPR -TPR و یک سری از چنین جفت‌هایی برای رسم منحنی‌های ROC استفاده می‌شوند که TPR به‌عنوان حساسیت[۲۱] و FPR –۱ به‌عنوان اختصاصیت[۲۲] شناخته می‌شوند. مقدار آستانهٔ زمانی روی خروجی تنظیم خواهد شد که تصمیم گرفته شود در این‌جا آلودگی وجود دارد یا ندارد. اگر احتمال، بزرگ‌تر از مقدار آستانه باشد طبقهٔ پیش‌بینی ۱ یا آلوده خواهد بود و اگر کمتر از مقدار آستانه باشد طبقهٔ پیش‌بینی صفر یا آلوده در نظر گرفته نمی‌شود. زمانی که حساسیت افزایش می‌یابد، مقدار اختصاصیت کاهش می‌یابد یا مقدار FPR افزایش می‌یابد. بنابراین آستانهٔ بهینه انتخاب می‌شود و برای طبقه‌بندی نقشه‌ها در تمایز نقاط آلوده و غیرآلوده استفاده می‌شود. همچنین مساحت زیر منحنی[۲۳] (AUC) در منحنی‌های ROC نیز به‌منزلهٔ یکی از روش‌های ارزیابی خطا به کار برده می‌شود که هرچه میزان AUC به یک نزدیک باشد، مدل از صحت بیشتری برخوردار است.

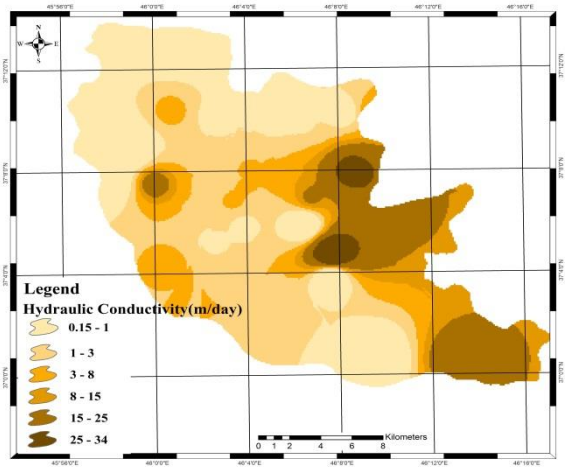
۳. بحث و نتایج

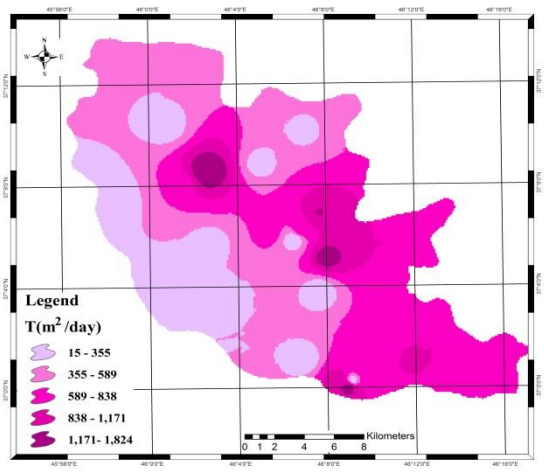
از بین ۲۳ متغیر استفاده‌شده در مدل، متغیرهای عمق سطح ایستابی یا به عبارت دیگر ضخامت زون وادوز، هدایت هیدرولیکی، فاصله از باغات انگور، گرادیان هیدرولیکی مطلق و قابلیت انتقال آبخوان می‌توانند رفتار نیترات در

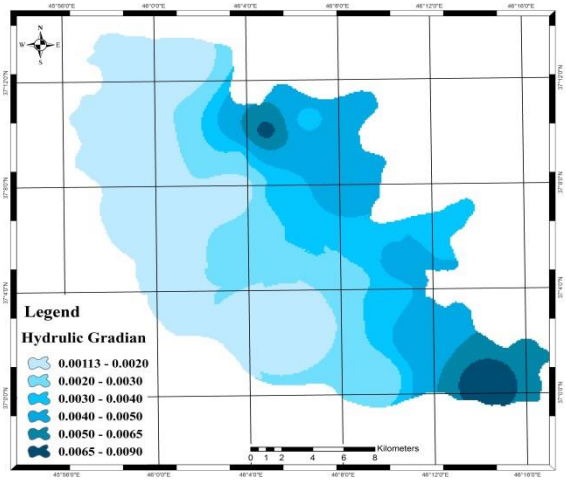
آبخوان دشت ملکان را با دقت بالایی توصیف کنند. هدایت هیدرولیکی در واقع شدت جریانی است که آب زیرزمینی تحت شیب هیدرولیکی محیط جریان پیدا می‌کند (Todd, 1980). هدایت هیدرولیکی عامل کنترل‌کنندهٔ حرکت و زمان ماندگاری مواد آلاینده، از نقطه‌ای که وارد سطح خاک می‌شود تا داخل سفره، است. به همین علت افزایش K باعث پتانسیل آلودگی بیشتر است. اطلاعات مربوط به هدایت هیدرولیکی از محاسبات آزمایش پمپاژ حاصل می‌شود و در مناطقی که آزمایش پمپاژ انجام نشده است، بر اساس نوع و بافت رسوبات تشکیل دهندهٔ آبخوان هدایت هیدرولیکی تخمین زده می‌شود. شکل ۷ لایه‌های رستری متغیرهای توضیحی استفاده‌شده در روش RF را نشان می‌دهد که در لایه‌های ایجادشده توزیع مکانی متغیرهای توضیحی مشاهده می‌شود. در منطقهٔ ملکان که یکی از قطب‌های تولید انگور کشور محسوب می‌شود، هر ساله برای افزایش بازدهی محصول باغات از کودهای حیوانی با مقدار بیش از اندازهٔ لازم استفاده می‌شود که این کودهای حیوانی به صورت مداوم میزان نیترات زیادی به سفرهٔ آب زیرزمینی وارده می‌کنند و پتانسیل آلودگی نیترات در منطقه را بالا می‌برند. همچنین، به دلیل وجود چاه‌های جذبی و نبود شبکهٔ فاضلاب در شهر ملکان، فاصله از شهرها نیز در گسترش آلودگی تأثیر زیادی دارد. جدول ۳ سهم هر متغیر توضیحی، همراه با میزان امتیاز پیش‌بینی‌شده برای هر کدام از متغیرها را در مدل نشان می‌دهد که به‌منزلهٔ یکی از خروجی‌های مدل است. برای محاسبهٔ قابلیت انتقال آبخوان دشت ملکان از نتایج آزمایش پمپاژ ۲۷ حلقه چاه استفاده شد که از طریق آب منطقه‌ای استان آذربایجان شرقی اندازه‌گیری شده است. با توجه به نقشهٔ قابلیت انتقال، بیشترین قابلیت انتقال در مرکز دشت است. ضریب ذخیرهٔ این دشت با روش بیلان جزء به جزء محاسبه شده است و با استفاده از روش مذکور، مقدار متوسط ضریب ذخیرهٔ آبخوان در حدود ۳/۲ درصد است.

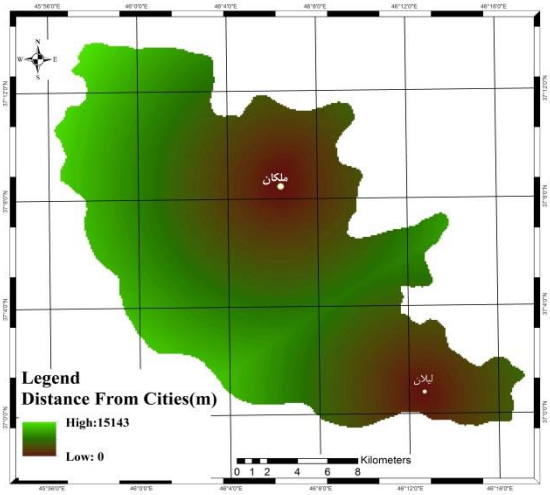
گرادیان هیدرولیکی مطلق که مسئول حرکت آب زیرزمینی و مهاجرت مادهٔ حل‌شونده در آن است در دشت ملکان در نقاط مختلف متفاوت است و به طور کلی در ورودی‌های دشت میانگین ۰/۰۰۶ و در خروجی‌ها نیز در حدود ۰/۰۰۱ است.

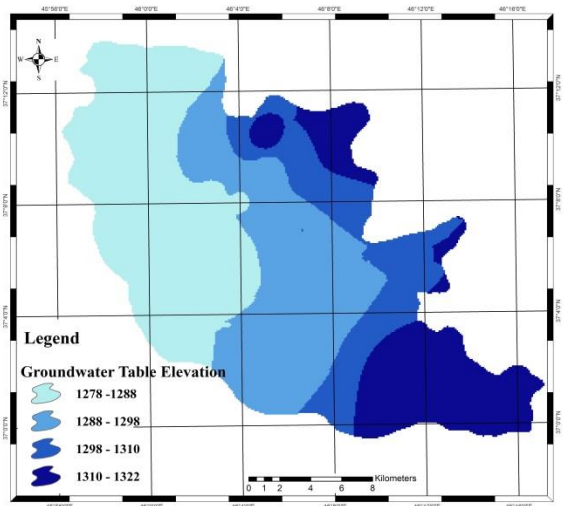
پیش‌بینی مدل RF با نرم‌افزار Salford Predictive Modeler انجام شد. برای انتخاب بهترین مدل پیش‌بینی غلظت نیترات در آبخوان دشت ملکان،RF با چهار زیرمجموعه از متغیرهای توضیحی به کار برده شد: مجموعهٔ داده‌های اول با همهٔ متغیرهای توضیحی، شامل خصوصیات ذاتی آبخوان، پارامترهای مربوط به کیفیت آب زیرزمینی و نیروهای محرک؛ مجموعهٔ داده‌های دوم با خصوصیات ذاتی آبخوان و پارامترهای کیفی؛ مجموعهٔ داده‌های سوم با نیروهای محرک و مجموعهٔ چهارم با پارامترهای روش دراستیک برای مدل‌سازی. در هر چهار مدل از غلظت نیترات به‌منزلهٔ متغیر پاسخ استفاده شد. وزن‌دهی پارامترها بر اساس تأثیر در احتمال آلودگی نیترات با کلاس‌بندی‌ای مناسب انجام گرفت. برای مثال هر چه به سمت باغات انگور یا شهرها نزدیک می‌شویم، به کلاسی که در آن فاصله از شهر یا باغات انگور کمتر است، به دلیل اینکه تأثیر بیشتری در آلودگی نیترات دارد، وزن بیشتری اختصاص داده شد و هرچه این فاصله بیشتر شود وزن کمتری به خود می‌گیرد. یا در بحث خواص هیدرولیکی آبخوان، هرچه مقدار قابلیت انتقال، هدایت هیدرولیکی و گرادیان هیدرولیکی بیشتر باشند بر اساس چارچوب هیدروژئولوژیکی وزن بیشتر به خود می‌گیرند و هرچه کمتر باشند به دلیل اینکه در احتمال آلودگی تأثیر کمتری دارند وزن کمتری به آن‌ها نسبت داده می‌شود. در نهایت بر اساس امتیاز پیش‌بینی‌شدهٔ مدل، همپوشانی همهٔ لایه‌ها انجام می‌گیرد و مکان‌هایی که احتمال تجاوز غلظت نیترات بیش از حد استاندارد جهانی وجود دارد شناسایی می‌شود.

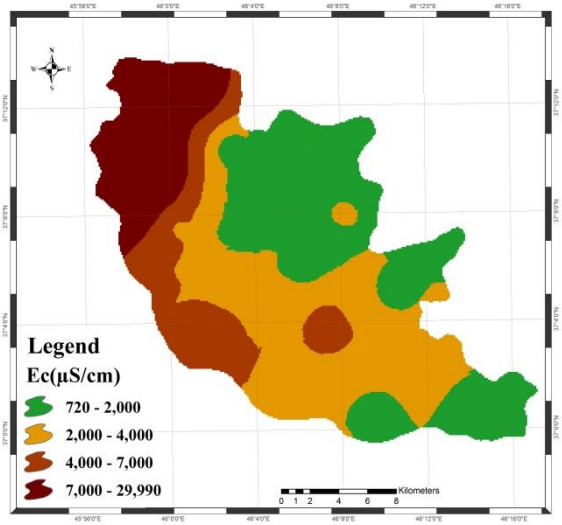

Legend
Hydraulic Conductivity(m/day)
0.15 - 1
1 - 3
3 - 8
8 - 15
15 - 25
25 - 34
Kilometers

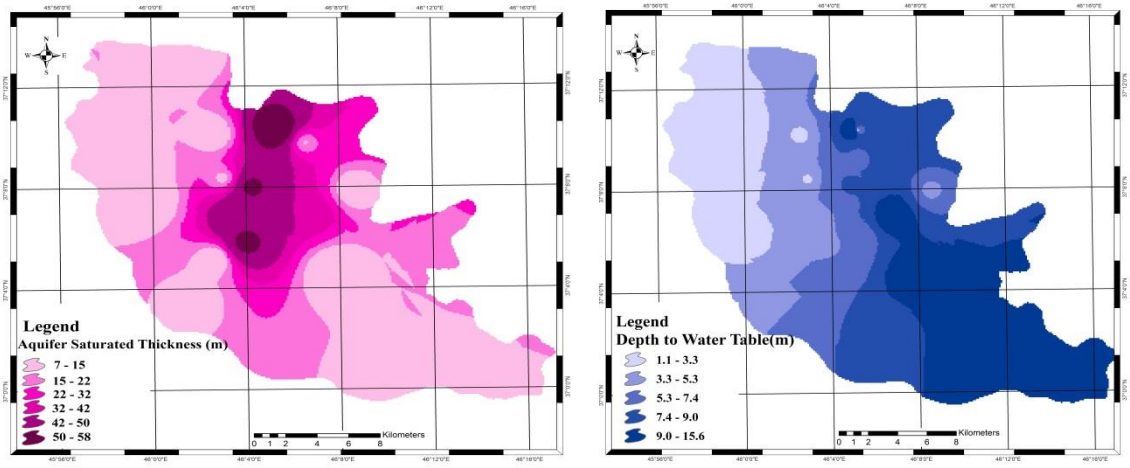

Legend
T(m2 /day)
15 - 355
355 - 589
589 - 838
838 - 1,171
1,171- 1,824
Kilometers


Legend
Hydrulic Gradian
0.00113 - 0.0020
0.0020 - 0.0030
0.0030 - 0.0040
0.0040 - 0.0050
0.0050 - 0.0065
0.0065 - 0.0090
Kilometers


ملکان
لیلان
Legend
Distance From Cities(m)
High:15143
Low: 0
Kilometers


Legend
Groundwater Table Elevation
1278 -1288
1288 - 1298
1298 - 1310
1310 - 1322
Kilometers


Legend
Ec(µS/cm)
720 - 2,000
2,000 - 4,000
4,000 - 7,000
7,000 - 29,990
Kilometers



شکل ۷. برخی از لایه‌های رستری مهم در پیش‌بینی مدل RF

جدول ۳. سهم هر متغیر توضیحی در مدل پیش‌بینی. A) تمام متغیرهای پیش‌بینی‌کننده، B) متغیرهای مربوط به خصوصیات ذاتی آبخوان و متغیرهای کیفی آب زیرزمینی، C) متغیرهای مربوط به نیروهای محرک، D) متغیرهای مربوط به پارامترهای مدل دراستیک

(A

سهم متغیر در پیش‌بینی	امتیاز	متغیر
	۱۰۰	عمق سطح ایستابی
	۷۰/۷۹	تراز آب زیرزمینی
	۴۴/۵۴	هدایت الکتریکی
	۳۳/۳۴	هدایت هیدرولیکی
	۳۰/۰۹	گرادیان هیدرولیکی
	۲۹/۲۸	قابلیت انتقال
	۲۷/۲۵	فاصله از شهرها
	۲۵/۲۲	فاصله از باغات انگور
	۲۵/۱۳	محیط خاک
	۱۸/۰۸	فاصله از مراکز صنعتی
	۱۶/۱۶	فاصله از محل‌های دفن زباله
	۱۶/۱۳	ضخامت اشباع آبخوان
	۱۶/۰۲	نرخ تغذیه
	۱۴/۳۲	ارتفاع کف آبخوان
	۱۳/۵۴	محیط آبخوان
	۶/۷۵	خاک
	۶/۷۳	فاصله از رودخانه‌ها
	۴/۲۷	شاخص پوشش گیاهی NDVI
	۲/۹۹	فاصله از کانال‌های آبیاری
	۱/۸۴	کاربری اراضی
	۱/۴۸	دما
	۰/۶۲	شیب سطح زمین
	۰/۲۵	اسیدیته

(B

سهم متغیر در پیش‌بینی	امتیاز	متغیر
	۱۰۰	عمق سطح ایستابی
	۸۲/۵۵	تراز آب زیرزمینی
	۴۶/۱۳	هدایت الکتریکی
	۳۶/۸۴	گرادیان هیدرولیکی
	۳۱/۰۳	هدایت هیدرولیکی
	۱۵/۲۳	محیط خاک
	۲۴/۸۴	قابلیت انتقال
	۱۴/۴۴	ارتفاع کف آبخوان
	۱۴/۳۹	محیط آبخوان
	۱۳/۴۸	ضخامت اشباع آبخوان
	۱۲/۰۵	خاک سطحی
	۱۱/۹۳	دما
	۳/۴۰	اسیدیته

(C

سهم متغیر در پیش‌بینی	امتیاز	متغیر
	۱۰۰	فاصله از شهرها
	۴۲/۸۴	فاصله از باغات انگور
	۳۹/۰۱	فاصله از رودخانه‌ها
	۳۸/۹۸	فاصله از مراکز صنعتی
	۱۳/۶۰	فاصله ازمحل‌های دفن زباله
	۱۱/۱۴	فاصله از کانال‌های آبیاری
	۳/۸۹	کاربری اراضی
	۱/۵۰	شاخص پوشش گیاهی NDVI

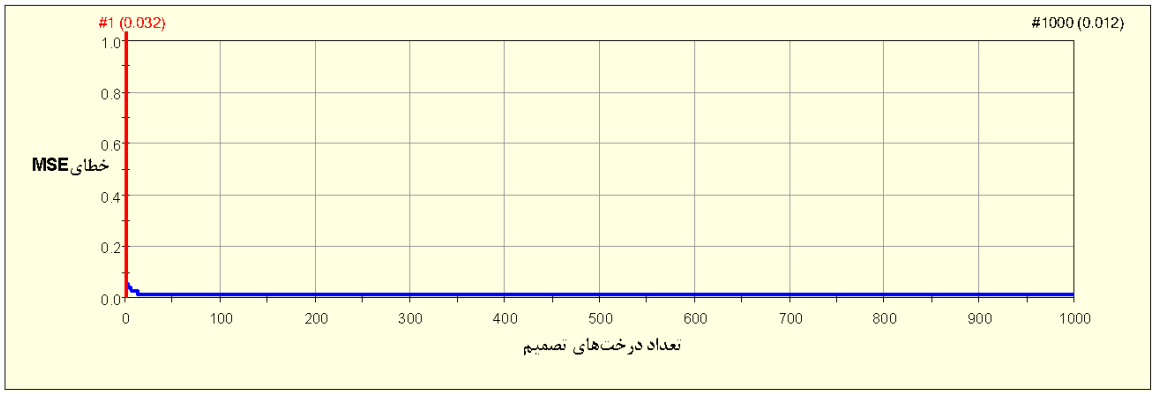
D)

متغیر	امتیاز	سهم متغیر در پیش‌بینی
عمق سطح ایستابی	۱۰۰	
هدایت هیدرولیکی	۴۵/۳۱	
محیط خاک	۳۰/۵۹	
محیط آبخوان	۱۹/۰۵	
نرخ تغذیه	۱۶/۴۳	
خاک سطحی	۶/۷۳	
شیب سطح زمین	۰/۷۸	

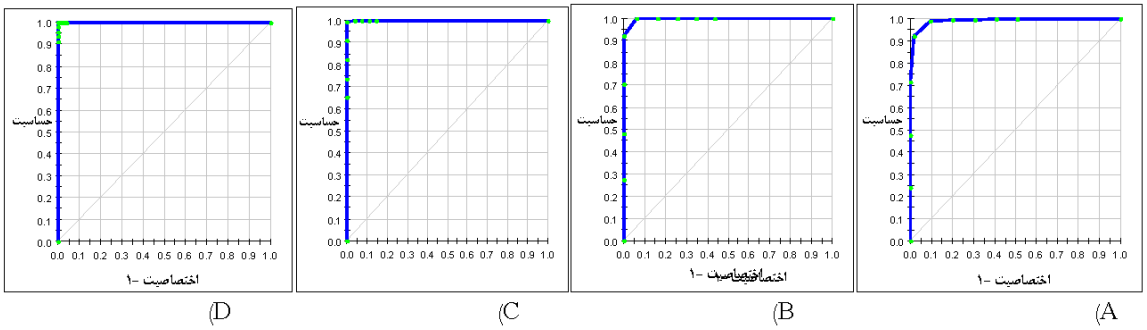
مدلی که کمترین مقدار MSE را داشته باشد به‌منزلهٔ بهترین و دقیق‌ترین مدل شناخته می‌شود. از بین چهار مدلی که با استفاده از روش جنگل تصادفی ارزیابی شدند، مدل نوع A که از همهٔ پارامترها، شامل متغیرهای خصوصیات ذاتی آبخوان، متغیرهای نیروهای محرک و متغیرهای دراستیک، استفاده کرد با داشتن کمترین مقدار MSE برابر با ۰/۰۱۲۴۵ و RMSE برابر با ۰/۱۱۱۵۷ به‌منزلهٔ دقیق‌ترین مدل انتخاب شد. در این مدل عمق سطح ایستابی نسبت به سایر پارامترها از اهمیت بیشتری در مدل پیش‌بینی برخوردار است و هرچه این عمق کمتر باشد پتانسیل آلودگی در منطقه بیشتر است. در دشت ملکان، با توجه به اینکه در قسمت‌های خروجی دشت به دلیل ریزدانه‌بودن رسوبات که باعث می‌شود آب در اثر نیروی کاپیلاری بالا بیاید و در عمق کمتر واقع شود، و به دلیل اینکه این مناطق هدایت هیدرولیکی، قابلیت انتقال و گرادیان هیدرولیکی کمتری دارند و از مناطق شهری، صنعتی و کشاورزی دورند، میزان آسیب‌پذیری به نیترات نسبت به سایر مناطق کمتر است و به تبع نزدیک‌بودن آب زیرزمینی به سطح زمین میزان هدایت الکتریکی در اثر تبخیر زیاد بالاست و در این مناطق میزان هدایت الکتریکی رابطه‌ای عکس با آلودگی نیترات دارد. در این مدل متغیرهای مربوط به pH، دما و شیب تأثیری در پیش‌بینی مدل ندارند، اما می‌توانند در برآورد خطای پیش‌بینی کمک کنند (جدول ۳، A). مدل نوع B که از متغیرهای مربوط به خصوصیات ذاتی آبخوان و پارامترهای مربوط به کیفیت آب زیرزمینی استفاده کرده است مدلی دقیق‌تر نسبت به مدل‌های C و D است و صحت نتایج آن از این دو روش بهتر است به طوری که MSE و RMSE آن به ترتیب برابر با ۰/۰۱۲۹۸ و ۰/۱۲۲۱۴ است و در این مدل نیز، عمق سطح ایستابی و تراز آب زیرزمینی متغیرهای مهمی در پیش‌بینی مدل محسوب می‌شوند و نقش مهمی در فرایند پیش‌بینی دارند (جدول ۳، B). در مدل نوع C که از متغیرهای مربوط به نیروهای محرک استفاده می‌کند به ترتیب، فاصله از مناطق شهری به دلیل نبود شبکهٔ فاضلاب این مناطق و فاصله از باغات انگور بیشترین اهمیت را در پیش‌بینی مدل دارند و شاخص NDVI اهمیتی در پیش‌بینی مدل ندارد. به طور کلی دقت این مدل با داشتن مقدار MSE برابر با ۰/۰۱۴۹۲و RMSE برابر ۰/۱۳۹۲ بیشتر از مدل پیش‌بینی‌شده با متغیرهای دراستیک است، اما به دلیل اینکه در این مدل خصوصیات آبخوان که عامل مهمی در انتقال آلاینده‌اند در نظر گرفته نشده است، نمی‌تواند قابل اعتمادتر باشد. در مدلی که از متغیرهای دراستیک استفاده شده است میزان خطا بیشتر از سه مدل ذکرشده است و مقدار MSE و RMSE آن به ترتیب برابر با ۰/۰۲۵۵۱ و ۰/۱۵۹۷۰ است. شکل ۸ میزان خطاهای حاصل از مدل A را که بهترین مدل پیش‌بینی است نشان می‌دهد.

انتخاب مدل پیش‌بینی بر اساس سه پارامتر دقت مدل پیش‌بینی، تعداد متغیرهای لازم و دردسترس‌بودن داده‌ها و دقت این داده‌ها پایدار خواهد بود. این مطالعه نشان می‌دهد که دقیق‌ترین مدل برای پیش‌بینی نیترات آبخوان دشت ملکان با استفاده از زیرمجموعه‌های مدل‌های A و B است و MSE تقریباً یکسانی دارند و مدل دراستیک بیشتر در مناطق با وسعت کمتر کاربرد دارد و هنگامی که در روش RF از این مدل استفاده می‌شود، دقت خیلی پایین‌تری نسبت به سایر روش‌ها دارد. شکل ۹ نتایج آنالیز ROC را برای هر سه مدل نشان می‌دهد که با رسم TPR (حساسیت) در مقابل FPR (اختصاصیت -۱) به دست می‌آید. همان طور که در منحنی‌های ROC مشاهده می‌شود

مدل A و B با داشتن بیشترین سطح زیر منحنی (AUC) بیشترین دقت را دارند و در مدل دراستیک میزان AUC کمتر و به تبع آن دقت مدل نیز نسبت به سایر مدل‌ها کمتر است و عملاً استفاده از پارامترهای دراستیک در روش جنگل تصادفی شکست‌خورده محسوب می‌شود. مقدار AUC چهار مدل A، B، C و D به ترتیب برابر با ۱/۰۰، ۰/۹۹۷، ۰/۹۹۶، ۰/۹۴۳ است و در نتیجه بهترین مدل، مدل نوع A است. البته خطای موجود در مدل نهایی به عوامل مختلفی بستگی دارد. از آن جمله می‌توان به تعداد محدود داده‌ها، ناهمگنی آبخوان، خطای ذاتی موجود در داده‌های ورودی و حتی داده‌های خروجی اشاره کرد که خطا در این داده‌ها پیش‌بینی مدل RF را با خطا مواجه می‌کند. افزایش دقت داده‌های ژئوفیزیکی و آزمایش پمپاژ در آبخوان‌ها می‌تواند در کاهش خطای موجود در مدل ارائه‌شده مؤثر باشد.



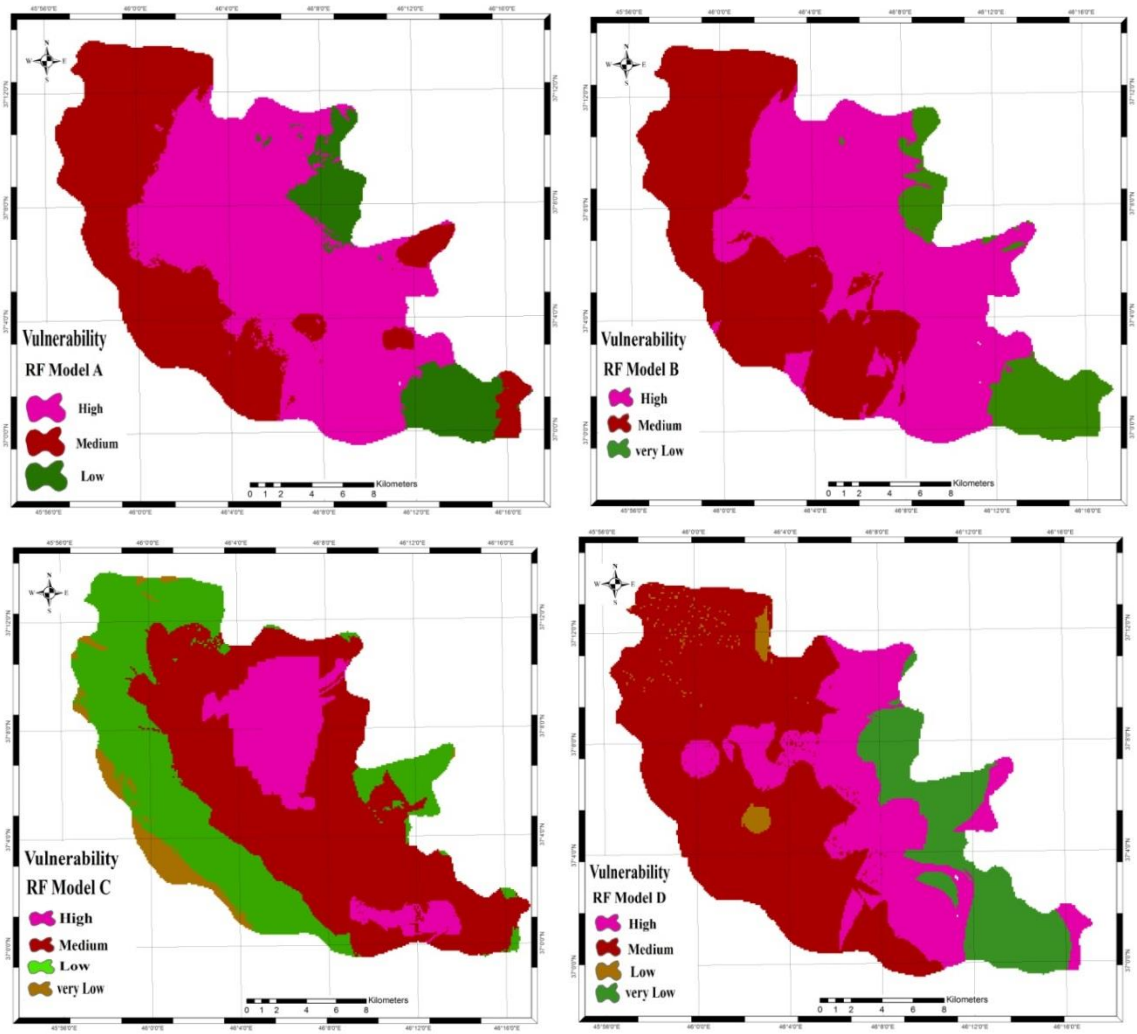
شکل ۸. خطای MSE مدل A و کاهش آن با افزایش درخت‌های تصمیم



شکل ۹. منحنی‌های ROC. A) تمام متغیرهای پیش‌بینی‌کننده، B) متغیرهای مربوط به خصوصیات ذاتی آبخوان و متغیرهای کیفی آب زیرزمینی، C) متغیرهای مربوط به نیروهای محرک، D) متغیرهای مربوط به پارامترهای مدل دراستیک

در روش‌هایی که تاکنون در بحث آسیب‌پذیری آب‌های زیرزمینی به کار رفته‌اند، علاوه بر داشتن داده‌های کمتر، از خود آلاینده‌ای مثل نیترات برای کالیبراسیون مدل استفاده نشده و بیشتر در صحت‌سنجی مدل به کار برده شده است، اما در روش RF از خود آلاینده به‌منزلهٔ بردار هدف مدل استفاده می‌شود و متغیرهایی که همبستگی بالایی با آلودگی نیترات دارند شناسایی می‌شوند. عملکرد بهتر روش جنگل تصادفی مربوط به توانایی آن در یادگیری

روابط غیرخطی بین نمونه‌های آلودهٔ نیترات و لایه‌های ورودی است. مزیت دیگر RF نسبت به روش‌های دیگر ماهیت غیرپارامتری آن است و نیاز به توزیع نرمال ندارد و علاوه بر این کارایی این روش برای نقاط دورافتاده بهتر از سایر روش‌هاست. شکل ۱۰ نقشه‌های آسیب‌پذیری حاصل از چهار نوع مدل را که در حقیقت بیان‌کنندهٔ مناطقی با احتمال تجاوز غلظت نیترات از حد مجاز است نشان می‌دهد. نتایج آسیب‌پذیری با مدل‌های صحیح از روش RF می‌تواند در مدیریت کیفی سفره‌های آب زیرزمینی بسیار کاربرد داشته باشد و با شناسایی مناطق حساس‌تر به آلودگی نقش بسزایی در این مسئله ایفا می‌کند. بر اساس نقشهٔ حاصل از مدل A، ۱۷ درصد از وسعت منطقهٔ مورد مطالعه در محدودهٔ آسیب‌پذیری کم، ۳۹ درصد در محدودهٔ آسیب‌پذیری متوسط، ۴۴ درصد از منطقه در محدودهٔ آسیب‌پذیری زیاد قرار می‌گیرد. با توجه به نقشهٔ آسیب‌پذیری، بیشترین درصد پتانسیل آلودگی دشت مربوط به کلاس آسیب‌پذیری بالاست که بیشتر در نواحی مرکزی منطقه مشاهده می‌شود. در مدل B، که نتایجی شبیه به نتایج مدل A داشت، ۴۲ درصد از منطقه مربوط به آسیب‌پذیری بالا، ۴۰ درصد مربوط به آسیب‌پذیری متوسط و ۱۸ درصد مربوط به آسیب‌پذیری پایین است.



شکل ۱۰. نقشه‌های آسیب‌پذیری حاصل از چهار نوع مدل. A) تمام متغیرهای پیش‌بینی‌کننده، B) متغیرهای مربوط به خصوصیات ذاتی آبخوان و متغیرهای کیفی آب زیرزمینی، C) متغیرهای مربوط به نیروهای محرک، D) متغیرهای مربوط به پارامترهای مدل دراستیک

نقشه‌های آسیب‌پذیری حاصل از مدل‌های C و D، که دقت آن‌ها پایین است و قابل اعتماد نیستند، به ترتیب ۱۵ و ۲۴ درصد از دشت مورد نظر در زون آسیب‌پذیری بالاتر، ۴۷ و ۵۱ درصد از منطقه در زون آسیب‌پذیری متوسط، ۲۹ و ۱۸ درصد منطقه در زون آسیب‌پذیری پایین و ۹ و ۷ درصد منطقه در زون با آسیب‌پذیری خیلی پایین قرار دارد.

همان طور که در نقشه‌های آسیب‌پذیری و نقشهٔ نیترات مشاهده می‌شود، بیشترین مقدار آلودگی و پتانسیل آلودگی به نیترات دشت ملکان در قسمت‌های شرقی و مرکزی این منطقه دیده می‌شود. با اینکه سطح آب‌زیرزمینی دشت ملکان در قسمت‌های غربی و شمال‌غربی در عمق کم واقع شده است که این عامل با توجه به پیش‌بینی RF مهمترین عامل در آلودگی نیترات محسوب می‌شود، اما به دلیل اینکه سایر عوامل مؤثر در آلودگی آب زیرزمینی به نیترات در قسمت‌های ذکرشدهٔ دشت مشاهده نمی‌شوند میزان آسیب‌پذیری در این قسمت‌ها در حدود متوسط است. برای مثال تراز آب زیرزمینی و هدایت هیدرولیکی، که در قسمت‌های شرقی و مرکزی آبخوان بیشترند، و هدایت الکتریکی نیز رابطهٔ عکس با میزان نیترات در این دشت دارند و در مکان‌هایی که شوری بیشتر است سایر عوامل آلودگی مؤثر در آلودگی کمتر است و از نظر جنس محیط غیراشباع و محیط آبخوان، مکان‌های مستعد به آلودگی که نفوذپذیری زیادی باید داشته باشند به سمت قسمت‌های مرکزی و شرقی درشت‌دانه‌ترند. گرادیان هیدرولیکی، قابلیت انتقال، فاصله از شهرها و فاصله از باغات انگور نیز از جمله متغیرهایی‌اند که مناطق مرکزی و شرقی آبخوان را در معرض آلودگی قرار می‌دهند. در بخش‌های کمتری از آبخوان دشت ملکان میزان آسیب‌پذیری آب زیرزمینی پایین است که علت آن عمق بیشتر سطح آب زیرزمینی در این مناطق است. سایر عوامل، مانند شیب زیاد نیز در پایین‌بودن میزان آسیب‌پذیری این مناطق مؤثر است و مطابقت نقشهٔ نیترات با نقشه‌های آسیب‌پذیری صحت نتایج مدل‌سازی را تأیید می‌کند. روش RF با شناسایی مناطق آسیب‌پذیر همچنین، با تشخیص و کنترل عوامل بیرونی تأثیرگذار در آلودگی آب زیرزمینی نقش بسزایی در مدیریت کیفی آب زیرزمینی ایفا می‌کند. با استفاده از نتایج این مدل می‌توان حریم کیفی منابع آب زیرزمینی را تعیین و مدیریت مناسبی برای کاربری اراضی مرتبط با سفرهٔ آبدار اعمال کرد.

۴. نتیجه‌گیری

در مطالعهٔ حاضر برای حفظ منابع آب زیرزمینی دشت ملکان از روش RF برای شناسایی مناطق آسیب‌پذیر نسبت به آلودگی نیترات استفاده شد. به دلیل اینکه روش RF از خود آلاینده مثل نیترات به‌منزلهٔ ورودی مدل در ترکیب با سایر پارامترها استفاده می‌کند و متغیرهایی را که همبستگی بالایی با آلودگی نیترات دارند شناسایی می‌کند، می‌تواند بر محدودیت سایر روش‌هایی که تاکنون در بحث آسیب‌پذیری آب‌های زیرزمینی استفاده شده‌اند غلبه کند. این مدل بر اساس چارچوبی هیدروژئولوژیکی برای آبخوان دشت ملکان به کار رفت. روش RF با داشتن مزایایی مثل یادگیری روابط غیرخطی، توانایی در مقابل داده‌های دورافتاده و حتی داده‌های ساختگی، برآورد خطای غیرمرتبط داخلی، اجرای هزاران دادهٔ ورودی بدون حذف یکی از آن‌ها و داشتن حساسیت کمتر در برابر گیرافتادن در مینیمم محلی و داده‌های پرت، به‌منزلهٔ روشی دقیق در مدل‌سازی پیش‌بینی آلودگی و آسیب‌پذیری شناخته شد. مدل RF نشان داد که عمق سطح ایستابی، تراز آب زیرزمینی، هدایت هیدرولیکی و هدایت الکتریکی به‌منزلهٔ متغیرهای مهم در پیش‌بینی آلودگی آب‌زیرزمینی مطرح‌اند و در حقیقت کنترل‌کنندهٔ آلودگی آبخوان‌اند. در این مطالعه که از چهار مدل برای پیش‌بینی آلودگی آبخوان به نیترات استفاده شد، مدل‌های A و B که به ترتیب از تمام متغیرها و متغیرهای ذاتی آبخوان استفاده کردند بیشترین دقت را داشتند. مدل‌های A و B، به ترتیب با RMSE مساوی با

۰/۱۱۱۵۷ و ۰/۱۲۲۱۴ ، حدود ۴۴ و ۴۲ درصد از منطقه را با محدودهٔ آسیب‌پذیری بالا پیش‌بینی کردند که در قسمت‌های مرکزی و شرقی آبخوان قرار دارند، در حالی که مدل‌های C و D، که از متغیرهای نیروهای محرک و دراستیک تشکیل شدند، به ترتیب با RMSE برابر ۰/۱۳۹۲ و ۰/۱۵۹۷۰، حدود ۱۵ و ۲۴ درصد از دشت مورد نظر را در محدودهٔ آسیب‌پذیری بالا قرار داده‌اند و با داشتن مقدار AUC کمتر ضعیف‌ترین نتایج را در مدل‌سازی پیش‌بینی آلودگی داشتند. مدل دقیق‌تر روش RF، با توجه به اهمیت آب زیرزمینی در منطقهٔ مورد مطالعه که برای مصارف مختلف کاربرد دارد، می‌تواند با شناسایی مناطق مستعد به آلوده‌شدن، منابع و عوامل مؤثر در آلودگی، برای مدیریت و نظارت صحیح آب‌های زیرزمینی استفاده شود.

یادداشت‌ها

1. Machine learning
2. Repeated Multiple Prediction
3. Land cover
4. Over fitting
5. Root Node
6. Terminal Root
7. Leaf Node
8. Pruning
9. Subset
10. Bagging
11. Out Of Bag
12. Decision-Maker
13. Feature Selection
14. Driving Force
15. Normalized Difference Vegetation Index
16. World Health Organization
17. Root Mean Square Error
18. True Positive Rate (TPR)
19. Receiver Operating Characteristic
20. False positive rate
21. Sensitivity
22. Specificity
23. Area under Curve

منابع

Antonakos, A. K. and Lambrakis, N.J. 2007. Development and testing of three hybrid methods for the assessment of aquifer vulnerability to nitrates, based on the drastic model, an example from NE Korinthia, Greece. Journal of Hydrology, Vol. 333(2), pp: 288–304.

Asghari Moghaddam, A., Fijani, E. and Nadiri, A. 2010. Groundwater Vulnerability Assessment Using GIS-Based DRASTIC Model in the Bazargan and Poldasht Plains. Journal of Environmental Studies, Vol. 35, pp: No. 52.

Bellman, R. 2003. Dynamic programming. Mineola, NY: Dover Publications 366 pp.

Booker, D.J. and Snelder, T. H. 2012. Comparing methods for estimating flow duration curves at ungauged sites. Journal of Hydrology 434–435, 78–94.

Boulesteix, A.L, Janitza, S. Kruppa, J, and König IR. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol.24 (2), pp: 493-507.

Breiman, L. 1996. Bagging predictors. Machine Learning, Vol. 24(2), pp: 40-123.

Breiman, L. 2001. Random Forests. Machine Learning, Vol. 45(1), pp: 5–32.

Chehata, N., Guo, L. and Mallet, C. 2009. Airborne lidar feature selection for urban classification using random forests. International Archives of the Photogrammetry, Journal of Remote Sensing and Spatial Information Sciences, Vol. 39, pp: 207-12.

Critto, A., Carlon, C. and Marcomini, A. 2003. Characterization of contaminated soil and groundwater surrounding an illegal landfill by principal component analysis and kriging. Journal of Environmental Pollution, Vol. 122(2), pp: 235–44.

Dixon, B.A. 2009. Case study using support vector machines, neural networks and logistic regression in a GIS to identify wells contaminated with nitrate-N. Journal of Hydrogeology, Vol. 17(6), pp: 1507–20.

Duda, R.O., Hart, P.E. and Stork, D.G. 2011. Pattern classification. 2nd. Edition. New York Efron B and Tibshirani R, 1993. In An introduction to the bootstrap. Vol. 57, pp: CRC press.

Emberger, L. 1952. Sur le quotient pluviothermique. C.R. Sciences, Vol. 234, pp: 2508-2511.

Fijani, E., Nadiri, A., Asghari Moghaddam, A., Tsai, F.T. C. and Dixon, B. 2013. Optimization of DRASTIC method by supervised committee machine artificial intelligence to assess groundwater vulnerability for Maragheh–Bonab plain aquifer, Iran. Journal of Hydrology, Vol. 503, pp: 89–100.

Friedl, M. A., Brodley, C. E. and Strahler, A. H. 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Trans Geoscience Remote Sensing, Vol. 37(2), pp: 969–77.

Gislason, PO. Benediktsson, JA. and Sveinsson, JR. 2004. Random forest classification of multisource remote sensing and geographic data. Journal of Geoscience and Remote Sensing Symposium, Vol. 2, pp: 1049-52.

Guo, L., Chehata, N., Mallet,C. and Boukir, S. 2011. Relevance of airborne lidar and multispectral imagedata for urban scene classification using Random Forests. ISPRS Journal of Photogram Remote Sensing, Vol. 66(1), pp: 56–66.

Guyon, I. and Elisseeff, A. 2003. An introduction to variable and feature selection. Journal of Machine Learning Reserch, Vol. 3, pp: 1157–82.

Ko, B., Gim, J. and Nam, J. 2011. Cell image classification based on ensemble features and random forest. Journal of Electronics Letters, Vol. 47, pp: 638-9.

Kotsiantis, S. and Pintelas, P. 2004. Combining bagging and boosting .International Journal of Computational Intelligence, Vol. 1(4), pp: 324–33.

Lehmann, P., and D. 2009. Evaporation and capillary coupling across vertical textural contrasts in porous media, journal of Physics, Vol. 80(4), pp: 18-46.

Nadiri, A., Fijani, E., Tsai,T.C. and Asghari Moghaddam, A.2013. Supervised Committee Machine with Artificial Intelligence for Prediction of Fluoride Concentration, Journal of Hydroinformatics. Vol. 15, pp: 1474–1490.

Pal, M. 2005. Random Forest classifier for remote sensing classification. International Journal of Remote Sensing, Vol. 26(1), pp: 217–22.

Pahlavan Rad, M.R., Toomanian, N., Khormali, F., Brungard, C., Komaki, C.B, and Bogaert, P. 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Journal of Geoderma, Vol. 232, pp: 97–106

Peters, J., Baets, B. D.,Verhoest, N. E. C., Samson, R., Degroeve, S. and Becker, P. D. 2007. Random Forests as a tool for ecohydrological distribution modelling. Journal of Ecol Model, Vol. 207(2–4), pp: 304–18.

Quinlan, J. R. 1993. C4.5 programs for machine learning. San Mateo, CA: Morgan Kaurmann 303 pp.

Quinlan, J.R. 1986. Induction of decision trees. Journal of Machine Learning, Vol. 1(1), pp: 81-106.

Rodriguez, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sánchez, J. P. 2012d. An assessment of the effectiveness of a Random Forest classifier for land-cover classification. ISPRS Journal of Photogramm Remote Sensing, Vol. 67, pp: 91-104.

Schapire, R. 1990. The strength of weak learnability. Journal of Machine learning, Vol. 5, pp: 197-227.

Thapinta, A. and Hudak, P. 2003. Use of geographic information systems for assessing groundwater pollution potential by pesticides in Central Thailand. International journal of Environmental, Vol. 29, pp: 87–93

Tilahun, K. and Merkel, B. J. 2010. Assessment of Groundwater Vulnerability to Pollution in Dire Dawa, Ethiopia using DRASTIC. Journal of Environmental Earth Sciences, Vol. 59, pp: 1485-1496.

Todd, D. K. 1980. Groundwater hydrology, John Wiley and Sons, New York.

Vrba, J. and Zoporozec, A. 1994. Guidebook on mapping groundwater vulnerability. International Contributions to Hydrogeology. 139 pp.

WHO (World Health Organization). 2009. Guideline for Drinking Water Quality.

Zabet, T. A. 2002. Evaluation of aquifer vulnerability to contaminant potential using DRASTIC method. Journal of Environmental Geology, Vol. 43(1-2), pp: 203-208.