نشریه پژوهش‌های حفاظت آب و خاک
جلد بیست و پنجم، شماره پنجم، ۱۳۹۷
http://jwsc.gau.ac.ir
DOI: 10.22069/jwsc.2018.12334.2691

مقایسه مدل‌های درخت تصمیم و یادگیری برپایه نمونه در برآورد هدایت هیدرولیکی اشباع خاک

مهنوش فرزادمهر[۱]، *مهدی دستورانی[۲] و عباس خاشعی‌سیوکی[۳]

[۱]دانش‌آموخته کارشناسی‌ارشد گروه علوم و مهندسی آب، دانشگاه بیرجند، [۲]استادیار گروه علوم و مهندسی آب، دانشگاه بیرجند،
[۳]دانشیار گروه علوم و مهندسی آب، دانشگاه بیرجند

تاریخ دریافت: ۹۷/۳/۸ ؛ تاریخ پذیرش: ۹۷/۶/۲۷

چکیده

سابقه و هدف: هدایت هیدرولیکی اشباع خاک یکی از مهم‌ترین خصوصیات هیدرولیکی خاک است که بر حرکت آب در خاک مؤثر است. شناخت این ویژگی می‌تواند به درک بسیاری از مشکلات زیست‌محیطی کمک کند. از طرفی اندازه‌گیری این ویژگی با روش‌های مستقیم مزرعه‌ای و آزمایشگاهی دشوار، زمان‌بر و هزینه‌بر است و استفاده از روش‌های جایگزینی را می‌طلبد که بتوان با صرف وقت، هزینه و زمان کم‌تری آن را از روی داده‌های زودیافت خاک تخمین زد. روش‌های ناپارامتریک از جمله روش‌های غیرمستقیم و نوین برآورد خصوصیات هیدرولیکی خاک از جمله هدایت هیدرولیکی اشباع می‌باشند. هدف از این پژوهش مقایسه روش درخت تصمیم و یک روش یادگیری برپایه نمونه (IBk) که یک رده‌بند با k همسایه نزدیک است در برآورد هدایت هیدرولیکی اشباع خاک، از روی خصوصیات زودیافت آن است.

مواد و روش‌ها: در این پژوهش، از مجموعه داده‌ای با اطلاعات خاک‌شناسی ۱۵۱ نمونه خاک که از منطقه‌ای در بجنورد گردآوری شده بود استفاده شد. خصوصیات زودیافت خاک شامل درصد شن، سیلت، رس، جرم مخصوص ظاهری، جرم مخصوص حقیقی، هدایت الکتریکی، درصد کربن آلی، درصد مواد خنثی‌شونده، رطوبت اشباع و اسیدیته بود. هدایت هیدرولیکی اشباع نمونه‌ها با استفاده از دستگاه نفوذسنج گلف اندازه‌گیری شده بود. برای تعیین مهم‌ترین پارامترها در پیش‌بینی و مدل‌سازی هدایت هیدرولیکی اشباع، از آزمون گاما استفاده شد. ترکیبات مختلف از پارامترهای موجود در بانک داده بر اساس مقدار گاما با یکدیگر مقایسه شدند و ترکیب بهینه برای مدل‌سازی معین شد. مدل‌سازی با استفاده از دو روش ناپارامتریک یعنی درخت تصمیم با بهره‌گیری از الگوریتم M5P و روش یادگیری برپایه نمونه با بهره‌گیری از الگوریتم IBk با استفاده از ترکیب بهینه پارامترها که کم‌ترین مقدار گاما را داشت صورت گرفت. برای بهبود عملکرد IBk دو نوع تابع وزن‌دهی فاصله استفاده شد. در آخر معیارهای ارزیابی مدل‌ها شامل ضریب تعیین (R^2)، جذر میانگین مربعات خطا (RMSE)، میانگین قدرمطلق خطا (MAE) و درصد میانگین قدرمطلق خطا (MAPE) محاسبه شدند.

* مسئول مکاتبه: mdastourani@birjand.ac.ir

یافته‌ها: ترکیب بهینه‌ای که از آزمون گاما به‌دست آمد برای مدل‌سازی هر دو روش استفاده شد. این ترکیب شامل پارامترهای درصد شن، سیلت، رس، درصد مواد خنثی‌شونده، هدایت الکتریکی و جرم مخصوص ظاهری خاک بود. مدل M5P، پارامتر جرم مخصوص ظاهری خاک را به‌عنوان مهم‌ترین متغیر دسته‌بندی‌کننده انتخاب کرد و سه رابطه خطی برای برآورد هدایت هیدرولیکی اشباع با توجه به مقدار جرم مخصوص ظاهری ایجاد کرد. معیارهای ارزیابی نشان دادند که این مدل با جذر میانگین مربعات خطای ۲۳/۸۹ سانتی‌متر بر روز و میانگین قدرمطلق خطای ۲۰/۵۰ درصد، دقت بالایی در پیش‌بینی هدایت هیدرولیکی اشباع نداشت. استفاده از دو نوع تابع وزن‌دهی تأثیری بر بهبود نتایج مدل IBk نداشتند. مدل IBk نیز با جذر میانگین مربعات خطای ۳۱/۲۳ سانتی‌متر بر روز و میانگین قدرمطلق خطای ۲۳/۲٤ درصد دقت بالایی نداشت.

نتیجه‌گیری: برای برآورد هدایت هیدرولیکی اشباع، درخت تصمیم مدل مناسب‌تری نسبت به مدل یادگیری برپایه نمونه بود، همچنین این مدل اطلاعاتی از ساختار خاک تحت بررسی نیز به‌دست داد.

واژه‌های کلیدی: آزمون گاما، الگوریتم IBk، الگوریتم M5P

مقدمه

خصوصیات فیزیکی خاک نقش تعیین‌کننده‌ای در درک و حل بسیاری از مشکلات زیست‌محیطی دارند. یکی از مهم‌ترین این خصوصیات هدایت هیدرولیکی اشباع است که نشان‌دهنده توانایی خاک در انتقال آب تحت شرایط اشباع می‌باشد. هدایت هیدرولیکی اشباع خاک (k_s) یک پارامتر مهم در تعیین عملیات آبیاری، طراحی زهکش‌ها، رواناب، تغذیه آب‌های زیرزمینی، شبیه‌سازی آبشویی و سایر فرآیندهای هیدرولوژیکی و کشاورزی است (۹). هدایت هیدرولیکی اشباع خاک نقش مهمی در نفوذ آب و ایجاد رواناب سطحی دارد.

برای اندازه‌گیری هدایت هیدرولیکی اشباع، روش‌های مختلف آزمایشگاهی و مزرعه‌ای استفاده می‌شوند که روش‌های آزمایشگاهی برای پژوهش مناسب‌ترند و روش‌های مزرعه‌ای بیش‌تر برای پروژه‌های زهکشی به‌کار می‌روند. هدایت هیدرولیکی اشباع تغییرات مکانی و زمانی بالایی دارد و در بین پارامترهای فیزیکی خاک دارای بیش‌ترین ضریب تغییرات است (۲۱). در صورتی‌که اندازه یا حجم نمونه برداشت‌شده افزایش یابد، نسبت به زمانی‌که حجم نمونه برداشت‌شده کم‌تر باشد، تخمین هدایت هیدرولیکی اشباع و تغییرات مکانی آن واقعی‌تر صورت می‌گیرد اما هزینه‌های بالای کارگری و مشکلات بیش‌تر نمونه‌برداری را در پی دارد (۱۷). با وجود تلاش‌های فراوانی که برای اندازه‌گیری هدایت هیدرولیکی اشباع در شرایط مزرعه انجام می‌گیرد هم‌چنان عواملی هستند که باعث ایجاد خطا در اندازه‌گیری این ویژگی می‌شوند از جمله این عوامل عبارتند از: حبس هوا در منافذ خاک هنگامی که نمونه خاک اشباع می‌شود، عدم اشباع کامل نمونه خاک هنگام استفاده از روش‌های اندازه‌گیری هدایت هیدرولیکی اشباع بالای سطح ایستابی و متفاوت بودن کیفیت آب مورد استفاده برای اشباع خاک با آب زیرزمینی همان محل که باعث تفاوت نتایج روش‌های اندازه‌گیری بین روش‌های بالای سطح ایستابی و زیر سطح ایستابی می‌شود (۱٦).

با توجه به دشواری، نیاز به زمان و هزینه بالای اندازه‌گیری هدایت هیدرولیکی اشباع با روش‌های

مستقیم، روش‌های غیرمستقیم به‌عنوان راهکاری برای حل نسبی این مشکلات ارائه شده‌اند. در روش‌های غیرمستقیم، از خصوصیات زودیافت خاک برای برآورد خصوصیات دیریافت آن استفاده می‌شود. یکی از این روش‌ها توابع انتقالی است که اولین بار توسط بوما در سال ۱۹۸۹ پیشنهاد و در مسائل مربوط به فیزیک خاک به‌کار گرفته شدند. وی توابع انتقالی را به‌صورت ترجمه داده‌هایی که موجود است به داده‌هایی که نیاز است بیان کرد (۱). پژوهشگران تاکنون توابع انتقالی مختلفی برای تخمین هدایت هیدرولیکی اشباع از روی خصوصیات زودیافت خاک ایجاد کرده‌اند. بعضی از این توابع نسبت به سایرین برآورد بهتری داشته در حالی‌که برخی مناسب نبودند، ترابی (۲۰۰٤) با مقایسه توابع انتقالی با روش‌های مستقیم نشان داد که مدل‌های ساکستون و همکاران (۱۹۸٦) با ورودی‌های شن و رس و نیز مدل جبرو (۱۹۹۲) با ورودی‌های سیلت، رس و جرم مخصوص ظاهری به‌دلیل عدم در نظر گرفتن سایر عوامل مؤثر بر هدایت هیدرولیکی اشباع مانند شوری خاک، مدل‌های مناسبی نیستند (۲٤). رسول‌زاده و همکاران (۲۰۱۲) با مقایسه توابع انتقالی رگرسیونی و نرم‌افزار روزتا (ROSETTA) و سویل پار (SOIL PAR) نشان دادند که مدل وستن و همکاران (۱۹۹۹) با ورودی‌های سیلت، جرم مخصوص ظاهری و ماده آلی و پس از آن مدل کاسبای و همکاران (۱۹۸٤) نتایج بهتری نسبت به دیگر توابع انتقالی داشته و عملکرد بهتر مدل وستن و همکاران را در استفاده از بانک داده قوی‌تر برای توسعه این مدل ذکر کرد (۲۲). شاپ و همکاران (۲۰۰۱) از نرم‌افزار روزتا که برپایه شبکه‌های عصبی است و از مدل معلم- ونگنوختن بهره می‌گیرد به پیش‌بینی هدایت هیدرولیکی اشباع و غیراشباع پرداختند آن‌ها نتیجه گرفتند که با افزایش پارامترهای ورودی به مدل، عملکرد توابع انتقالی در پیش‌بینی k_s قابل‌قبول بود (۲۳). ویژگی مشترک توابع انتقالی این است که همه آن‌ها بر اساس روش‌های پارامتریک هستند، یعنی معادله‌هایی از پیش تعیین شده با تعدادی پارامتر معین هستند. از معایب روش‌های پارامتریک می‌توان به این موارد اشاره نمود: تشخیص معادله صحیح همیشه آسان نیست، در صورتی‌که داده جدید فراهم شود معادلات باید بازسازی شوند و کاربران قادر نخواهند بود که به سادگی هر گونه داده اضافی را برای بهبود عملکرد در منطقه خود و برای خصوصیات ویژه خاک خود استفاده کنند (۲۰). استفاده از روش‌های ناپارامتریک یک راه‌حل جایگزین است. یعنی روش‌هایی که فرضیاتی راجع به روابط متغیرها ایجاد نمی‌کنند و از توابع از پیش تعیین‌شده‌ای برای پیش‌بینی متغیر موردنظر استفاده نمی‌کنند. از روش‌های ناپارامتریک که از الگوریتم‌های ناپارامتریک استفاده می‌کنند می‌توان به روش درخت تصمیم و k- نزدیک‌ترین همسایه (k-NN) اشاره نمود. یکی از موارد استفاده از روش‌های ناپارامتریک برای پیش‌بینی خصوصیات هیدرولیکی خاک پژوهش نمس و همکاران (۲۰۰٦) است که k-NN را برای تخمین نگه‌داشت آب در دو پتانسیل ماتریک ۳۳- و ۱۵۰۰- کیلوپاسکال با استفاده از داده‌های زودیافت خاک استفاده کردند و برتری این روش را در امکان استفاده از داده‌های محلی بیان نمودند (۲۰). حق‌وردی و همکاران (۲۰۱۰) برای برآورد میزان رطوبت در دو نقطه ظرفیت زراعی و پژمردگی دائم از این روش استفاده کردند (۸). جلالی و همایی (۲۰۱۱) k-NN را برای برآورد جرم مخصوص ظاهری خاک با استفاده از سایر متغیرهای کمکی خاک به‌کار بردند (۱۰). خاشعی‌سیوکی و همکاران (۲۰۱۵) با استفاده از داده‌های زودیافت خاک k-NN و سیستم‌های شبکه عصبی مصنوعی را برای برآورد هدایت هیدرولیکی اشباع به‌کار بردند و نشان دادند که هر دو مدل دارای توانایی خوبی در تخمین هدایت هیدرولیکی اشباع هستند و k-NN در

مقایسه با شبکه عصبی مصنوعی دارای توانایی بالاتری در تخمین هدایت هیدرولیکی به‌ازای پارامترهای ورودی کم‌تر می‌باشد (۱٤). در بیش‌تر پژوهش‌ها برای برآورد هدایت هیدرولیکی اشباع، روش‌های نزدیک‌ترین همسایه استفاده شده‌اند که در کاربرد این روش‌ها متغیرها باید نرمال شوند (۲۰). الگوریتم برپایه نمونه IB1 از نزدیک‌ترین همسایه‌ها ایجاد شده‌اند و دامنه متغیرها را نرمال می‌کنند (۲). اما تاکنون پژوهشی با استفاده از این روش و استفاده از توابع مختلف وزن‌دهی فاصله برای بهبود عملکرد آن در نرم‌افزار WEKA صورت نگرفته است. درخت تصمیم یکی دیگر از روش‌های داده‌کاوی است که در علوم خاک استفاده شده است. مونکادا و همکاران (۲۰۱٤) مدل‌های درخت تصمیم را برای ارزیابی کیفیت خاک و تخمین هدایت هیدرولیکی اشباع در دو خاک گرمسیری و معتدل استفاده کردند. آن‌ها از خصوصیات مورفولوژیکی خاک علاوه بر خصوصیات شیمیایی و فیزیکی آن استفاده کردند و مشاهده کردند که کاربرد خصوصیات مورفولوژیکی در کنار سایر خصوصیات خاک تخمین k_s بهبود می‌یابد و درختی با یک ساختار ساده‌تر ایجاد می‌شود (۱۹). یکی از مهم‌ترین مراحل در مدل‌سازی تعیین یک ترکیب مناسب از پارامترهای ورودی به‌منظور افزایش کارایی مدل است. اگر n پارامتر بر پدیده‌ای مؤثر باشند، تعداد $2^n - 1$ ترکیب از این پارامترها به وجود می‌آید که بررسی هر یک از این ترکیبات در مدل‌سازی زمان‌بر است. با استفاده از آزمون گاما مهم‌ترین پارامترها و بهترین ترکیبی که منجر به کم‌ترین خطای مدل‌سازی می‌شود را می‌توان تعیین کرد. در الگوریتم‌های ناپارامتریک داده‌کاوی، با وجود این‌که می‌توان از تعداد زیادی پارامتر برای پیش‌بینی متغیر وابسته استفاده کرد، اما بهتر است که برای کاهش عملیات اندازه‌گیری پارامترهای زودیافت خاک پارامترهای مهم‌تر تعیین شود. مقدم‌نیا و همکاران (۲۰۰۹) از آزمون گاما برای انتخاب ترکیب بهینه از پارامترهای ورودی برای مدل‌سازی تبخیر با استفاده از شبکه عصبی مصنوعی و سیستم نروفازی استفاده کردند (۱۸). قبائی‌سوق و همکاران (۲۰۱۰) از آزمون گاما به‌عنوان روشی برای پیش‌پردازش داده‌ها استفاده کردند ایشان با استفاده از آزمون گاما، مهم‌ترین پارامترهای مؤثر بر تبخیر و تعرق روزانه، ترکیب بهینه از پارامترها و تعداد داده‌های لازم برای آموزش و آزمون شبکه عصبی مصنوعی را مشخص کردند (۷). هدف از این پژوهش مقایسه دو مدل ناپارامتریک یعنی درخت تصمیم با بهره‌گیری از الگوریتم M5P و یادگیری برپایه نمونه با بهره‌گیری از الگوریتم IBk که یک رده‌بند با k همسایه نزدیک است و براساس الگوریتم IB1 کار می‌کند، برای پیش‌بینی هدایت هیدرولیکی اشباع با استفاده از نرم‌افزار WEKA می‌باشد. ترکیب بهینه از پارامترهای ورودی با استفاده از آزمون گاما مشخص و برای مدل‌سازی استفاده شد.

مواد و روش‌ها

در این پژوهش از سری داده‌های با اطلاعات خاکشناسی ۱۵۱ نمونه خاک استفاده شد (۱٤). این مجموعه داده از منطقه دشت دامنه‌ای قره‌میدان واقع در ۷۰ کیلومتری شمال‌غرب بجنورد تهیه شده (۱٤). هدایت هیدرولیکی اشباع با استفاده از دستگاه نفوذسنج گلف اندازه‌گیری شد، جرم ویژه ظاهری نمونه‌ها به روش کلوخه، جرم ویژه حقیقی از طریق پیکنومتر، فراوانی نسبی انداز ذرات به روش هیدرومتری تعیین شد (به نقل از ۱٤). خصوصیات زودیافت خاک و توصیف آماری آن‌ها در جدول ۱ آورده شده است. همچنین برای هر نمونه مختصات جغرافیایی آن نیز در دسترس بود که این ویژگی نیز برای مدل‌سازی استفاده شد.

جدول ۱- توصیف آماری خصوصیات خاک مورد مطالعه.

Table 1. Statistical description of the studied soil.

ضریب تغییرات C.V (%)	میانگین Mean	بیشینه Maximum	کمینه Minimum	دامنه Range	واحد Unit	ویژگی Property
38.99	0.21	0.7	0.037	0.663	$g\ g^{-1}$	شن Sand
10.77	0.49	0.62	0.184	0.436	$g\ g^{-1}$	سیلت Silt
15.04	0.3	0.436	0.116	0.32	$g\ g^{-1}$	رس Clay
24.91	0.89	1.91	0.21	1.7	%	کربن آلی OC
24.38	0.22	47.75	5.75	0.42	%	مواد خنثی‌شونده TNV
28.03	1.74	3.81	0.27	3.54	$dS.m^{-1}$	هدایت الکتریکی EC
4.02	7.36	8.33	6.44	1.89	$-Log\ [H^+]$	اسیدیته pH
4.82	50.5	56	43	13	$m^3\ m^{-3}$	رطوبت اشباع Saturation moisture
3.43	2.65	2.78	2.19	0.59	$g\ cm^{-3}$	جرم مخصوص حقیقی Pd
5.59	1.53	1.75	1.26	0.49	$g\ cm^{-3}$	جرم مخصوص ظاهری Bd
115.99	17.12	194.62	1.39	193.22	$cm\ d^{-1}$	هدایت هیدرولیکی اشباع K_s

آزمون گاما: آزمون گاما برای تخمین میزان خطای موجود در داده‌ها است که آن‌را به‌طور مستقیم از روی داده‌ها محاسبه می‌کند و هیچ فرضیاتی راجع به روابط پارامتریک معادلات حاکم بر سیستم موردنظر ایجاد نمی‌کند (۸). با استفاده از آزمون گاما ترکیب بهینه از متغیرهای ورودی برای هر گونه مدل‌سازی غیرخطی را می‌توان به‌دست آورد. اساس کار این آزمون بر پایه نزدیک‌ترین همسایگی‌ها است (٦). به‌عبارتی چنان‌چه دو نقطه از فضای ورودی (x_1 و x_2) به قدر کافی به یکدیگر نزدیک باشند، خروجی متناظر با آن‌ها (y_1 و y_2) نیز باید به یکدیگر نزدیک باشند. چنان‌چه خروجی‌ها دارای قرابت نباشند، این تفاوت به‌عنوان خطا در نظر گرفته می‌شود (۱۲). رابطه ۱ رابطه نهفته در سیستم تحت بررسی را نشان می‌دهد:

$$y = f(x_1 \ \dots \ x_m) + r \qquad (۱)$$

که در آن، y متغیر خروجی، x متغیر ورودی، m تعداد پارامترها، f نشان‌دهنده یک تابع هموار[1] و r نشان‌دهنده یک متغیر تصادفی که نماینده خطا می‌باشد

1- Smooth function

است. آماره گاما آن بخش از واریانس خطا یا var(r) است که توسط یک مدل هموار نمی‌تواند برآورد گردد. هرچه گاما به صفر نزدیک‌تر باشد، نشان‌گر آن است که محدودیتی برای ساخت یک مدل هموار وجود ندارد. نسبت v، خطای استاندارد و شیب خط رگرسیونی اطلاعات مفید دیگری هستند که از این آزمون می‌توان به‌دست آورد. هرچه نسبت v به صفر نزدیک‌تر باشد یعنی مدل قابلیت پیش‌بینی بالایی از روی داده‌ها برای یک خروجی معین دارد. شیب خط رگرسیونی اطلاعاتی راجع به پیچیدگی مدل می‌دهد. خطای استاندارد میزان درستی رگرسیون خطی در آماره گاما را بیان می‌کند و اگر به صفر نزدیک شود مقدار گاما مطمئن‌تر خواهد بود (۱۱). در این پژوهش از نرم‌افزار وین گاما (WIN GAMMA) که توسط دانشگاه کاردیف طراحی شده برای محاسبات مربوط به آزمون گاما استفاده شد. داده‌ها قبل از استفاده در آزمون گاما، مدل M5P و IBk، طبق رابطه ۲ بین دو عدد ۰/۱ و ۰/۹ نرمال و استاندارد شدند (۱٤). داده‌ها پس از شبیه‌سازی به مقادیر اولیه برگشتند.

$$x_i = 0.8\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right) + 0.1 \qquad (۲)$$

که در آن، x_i مقدار استاندارد شده، x مقدار واقعی و $x_{\min}$ و $x_{\max}$ به‌ترتیب مقادیر حداقل و حداکثر داده‌ها می‌باشد.

درخت تصمیم: درخت تصمیم برای انجام پیش‌بینی، ساختاری مشابه درخت ایجاد می‌کند به این صورت که ابتدا کار خود را با استفاده از تمام نمونه‌های آموزشی شروع می‌کند و متغیری که بهترین دسته‌بندی را انجام می‌دهد انتخاب می‌کند و زیرمجموعه‌هایی تشکیل می‌دهد. شاخه‌های درخت نتیجه آزمونی است که در هر مرحله توسط الگوریتم بر روی گره‌های میانی صورت می‌گیرد. پیش‌بینی‌ها نیز در برگ‌های درخت ظاهر می‌شوند (۵). مدل درختی M5P قابلیت پیش‌بینی متغیرهای پیوسته عددی از روی صفات عددی را دارد و نتایج پیش‌بینی‌شده به‌صورت مدل‌های رگرسیونی خطی چندمتغیره در برگ‌های درخت ظاهر می‌شوند (۲۵). معیار تقسیم در یک گره بر اساس انتخاب انحراف معیار مقادیر خروجی که به آن گره می‌رسند به‌عنوان معیاری از خطا است. با آزمودن هر صفت (پارامتر) در گره کاهش مورد انتظار در خطا محاسبه می‌شود. کاهش انحراف معیار با رابطه ۳ محاسبه می‌شود (۲۵).

$$SDR = \frac{m}{|T|} \times \beta(i) \times \left[sd(T) - \sum_{j \in (L,R)} \frac{|T_j|}{|T|} \times sd(T_j) \right] \qquad (۳)$$

که در آن، SDR کاهش انحراف معیار است. T نشان‌دهنده سری نمونه‌هایی است که به گره می‌رسند، m تعداد نمونه‌هایی است که برای این صفت مقادیر گم‌شده ندارند، $\beta(i)$ یک عامل اصلاحی است و T_L و T_R مجموعه‌هایی هستند که از تقسیم بر روی این صفت به وجود می‌آیند. هرس[1] درخت به معنای حذف گره‌های اضافی برای جلوگیری از بیش‌برازش درخت به داده‌های آموزشی است. مرحله آخر ساخت مدل‌های درختی هموارسازی[2] است که برای جبران ناپیوستگی‌هایی که به ناچار میان مدل‌های خطی هم‌جوار در برگ‌های درخت هرس شده اتفاق می‌افتد صورت می‌گیرد (۲۵). در این پژوهش از الگوریتم M5P نرم‌افزار WEKA 3.8 برای مدل‌سازی استفاده شد. داده‌ها به دو بخش تقسیم شدند، ۷۵٪ برای آموزش مدل (۱۱۳ نمونه) و ۲۵٪ برای آزمون مدل (۳۸ نمونه) استفاده شدند.

روش یادگیری برپایه نمونه: روش‌های یادگیری برپایه نمونه[3] روش‌های ناپارامتریکی هستند که از

1- Pruning
2- Smoothing
3- Instance based learning methods

روابط از پیش تعیین‌شده برای پیش‌بینی متغیر مجهول استفاده نمی‌کنند و به جای آن از نمونه‌های مشخصی برای پیش‌بینی متغیر استفاده می‌کنند (۲). بعضی از الگوریتم‌های این روش شامل IB1، IB2 و IB3 هستند (۲). این روش‌ها فرض می‌کنند که نمونه‌های مشابه دارای دسته‌بندی مشابه هستند. بنابراین دسته‌بندی نمونه‌های جدید، براساس نمونه‌هایی که بیش‌ترین شباهت را با آن نمونه دارند انجام می‌شود. همچنین بدون داشتن هیچ اطلاعاتی از متغیرها، فرض می‌کنند که متغیرها دارای وزن یکسان در تابع تشابه هستند. روش‌های یادگیری برپایه نمونه از روش‌های نزدیک‌ترین همسایه توسعه یافته‌اند. الگوریتم IB1 ساده‌ترین این الگوریتم‌ها و مشابه الگوریتم‌های نزدیک‌ترین همسایه هستند با این تفاوت که IB1 دامنه متغیرها را نرمال می‌کند و روشی ساده در برخورد با مقادیر گمشده دارد (۲). در این روش، با استفاده از تعداد اندکی از نمونه‌های مشابه (همسایه‌های نزدیک)، فرض می‌شود که تابع موردنظر به‌صورت محلی خطی است و با ایجاد توابع تکه‌ای و خطی سعی در پیدا کردن تابع ناشناخته دارند. در این پژوهش از IBk که براساس IB1 کار می‌کند برای پیش‌بینی نمونه هدف استفاده شده است. IBK یک رده‌بند با k همسایه نزدیک است که می‌توان از k تعداد همسایه نزدیک و توابع وزن‌دهی فاصله استفاده کرد. به‌طور معمول فاصله اقلیدسی به‌عنوان تابع فاصله استفاده می‌شود (۲۰). در این پژوهش نیز از فاصله اقلیدسی برای محاسبه فاصله نمونه خاک هدف (آزمایشی) و نمونه‌های خاک مرجع (آموزشی) استفاده شد پارامتر بعدی که باید برای مدل‌سازی با IBk مشخص شود تعداد نزدیک‌ترین همسایه‌ها (k) است که به تعداد نمونه‌های خاک مرجع بستگی دارد. برای این پژوهش بیش‌ترین مقدار برای k برابر با ریشه دوم تعداد نمونه‌ها (۱۵) و کم‌ترین مقدار آن برابر یک در نظر گرفته شد و با استفاده از ارزیابی تقاطعی یکی بیرون[1] مقدار مناسب k به‌دست آمد. در این روش داده‌ها به چند بخش تقسیم می‌شوند و آنالیز بر روی یک زیرمجموعه انجام و بر روی مجموعه‌های دیگر ارزیابی می‌گردد. ارزیابی تقاطعی یکی بیرون همان ارزیابی تقاطعی n بخشی[2] است، با این تفاوت که n برابر با تعداد نمونه‌های بانک داده است. در این پژوهش از دو نوع تابع وزن‌دهی فاصله شامل وزن‌دهی معکوس و خطی برای بهبود عملکرد IBk استفاده شد که به‌ترتیب در رابطه‌های ٤ و ٥ آمده است و سپس نتایج آن‌ها با IBk بدون وزن‌دهی مقایسه شد.

$$w_i = \frac{1}{d_i} \qquad (٤)$$

$$w_i = 1 - d_i \qquad (٥)$$

که در آن، w_i وزن نمونه خاک است. d_i فاصله اقلیدسی نمونه i تا خاک هدف است. به این صورت همسایه‌های نزدیک‌تر وزن بیش‌تری به خود اختصاص می‌دهند و تأثیر بیش‌تری در پیش‌بینی متغیر هدف خواهند داشت. در مرحله آخر با محاسبه میانگین وزنی هدایت هیدرولیکی اشباع شبیه‌ترین نمونه‌ها می‌توان هدایت هیدرولیکی اشباع نمونه هدف را به‌دست آورد. برای مدل‌سازی با استفاده از الگوریتم IBk، داده‌ها به دو بخش آموزش و آزمون تقسیم شدند.

معیارهای ارزیابی مدل: معیارهای آماری ریشه متوسط خطای مربعات (RMSE)، ضریب تعیین (R^2)، متوسط قدرمطلق خطا (MAE) و درصد میانگین مطلق خطا (MAPE) برای ارزیابی کارایی مدل‌ها طبق رابطه‌های ٦ تا ٩ محاسبه شدند:

1- Leave one-out cross validation
2- N- fold cross validation

$$RMSE = \sqrt{\frac{(\sum (K_{Si} - K_{Oi})^2)}{N}} \quad (۶)$$

$$R^2 = \frac{(\sum_{i=1}^{N}(K_{Oi} - \overline{K}_O)(K_{Si} - \overline{K}_S))^2}{(\sum_{i=1}^{N}(K_{Oi} - \overline{K}_O)^2 \sum_{i=1}^{N}(K_{Si} - \overline{K}_S)^2)} \quad (۷)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|k_{Si} - k_{Oi}| \quad (۸)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{k_{Oi} - k_{Si}}{k_{Oi}}\right| \times 100 \quad (۹)$$

که در آن‌ها، N برابر تعداد کل داده‌ها، K_{si} هدایت هیدرولیکی اشباع پیش‌بینی‌شده، K_{oi} هدایت هیدرولیکی اشباع اندازه‌گیری شده، $\overline{K}_O$ میانگین مقادیر هدایت هیدرولیکی اشباع اندازه‌گیری‌شده، $\overline{K}_S$ میانگین مقادیر هدایت هیدرولیکی اشباع پیش‌بینی شده است.

نتایج و بحث

انتخاب پارامترهای ورودی با آزمون گاما: برای مشخص کردن اهمیت پارامترهای ورودی و ترکیب بهینه برای مدل‌سازی، دو مرحله انجام شد. ابتدا آزمون گاما برای ترکیبی که همه پارامترهای مؤثر بر هدایت هیدرولیکی اشباع که در بانک داده موجود بود محاسبه گردید. این ترکیب شامل یازده پارامتر درصد شن، درصد سیلت، درصد رس، درصد کربن آلی، درصد مواد خنثی‌شونده، هدایت الکتریکی، اسیدیته، رطوبت اشباع، جرم مخصوص حقیقی، جرم مخصوص ظاهری و طول و عرض جغرافیایی بود. در گام بعدی یکی از پارامترها به تنهایی از ترکیب اولیه که شامل همه یازده پارامتر بود حذف شد و آماره گاما برای ترکیب جدید که دارای ده پارامتر باقی‌مانده است محاسبه شد. سپس پارامتر حذف‌شده به مجموعه اولیه بازگردانده شد و یک پارامتر دیگر حذف شد و مجدداً آماره گاما برای این ترکیب جدید محاسبه گردید. این فرآیند تا جایی ادامه داشت که هر یک از پارامترها یک بار از ترکیب اولیه حذف شده باشند و آماره گاما برای ترکیب جدید محاسبه شده باشد. نتایج در جدول ۲ آورده شده است.

مقدار گاما برای ترکیب شماره یک که همه پارامترهای ورودی برای مدل‌سازی در نظر گرفته شده است ۰/۰۰۰۲۰ به‌دست آمده است. حذف هریک از پارامترها از این ترکیب باعث تغییر در گاما شد. با توجه به نتایج جدول ۲، بیش‌ترین افزایش مقدار گاما نسبت به ترکیب شماره یک هنگامی صورت گرفت که فقط جرم مخصوص ظاهری خاک حذف شد (یعنی ترکیب شماره دوازده) و این امر نشان می‌دهد که برای فرآیند مدل‌سازی هدایت هیدرولیکی اشباع، پارامتر جرم مخصوص ظاهری خاک بیش‌ترین اهمیت را نسبت به سایر پارامترها دارد. حذف هر یک از پارامترهای شن، سیلت، مواد خنثی‌شونده، هدایت الکتریکی و جرم مخصوص حقیقی باعث افزایش در گاما شده است و می‌توان نتیجه گرفت بعد از جرم مخصوص ظاهری، این پارامترها در مرتبه دوم اهمیت قرار دارند. حذف هر یک از پارامترهای درصد رس، کربن آلی و اسیدیته در تغییر گاما بی‌تأثیر بوده است. حذف طول و عرض جغرافیایی و رطوبت اشباع باعث کاهش گاما نسبت به ترکیب یک شده است. پس می‌توان گفت که وجود طول و عرض جغرافیایی و رطوبت اشباع در ترکیب با دیگر پارامترها موجب خطای بیش‌تری در مدل‌سازی می‌شود و وجود آن‌ها در ترکیب بهینه الزامی نیست. در این پژوهش، با یازده پارامتر مؤثر بر هدایت هیدرولیکی اشباع تعداد ۲۰۴۷ ترکیب به وجود آمد (دو پارامتر طول و عرض جغرافیایی با هم در نظر گرفته شد) که با استفاده از آزمون گاما، مقدار گاما برای این ترکیبات به‌دست آمد. پنج ترکیبی که کم‌ترین گاما را داشتند به‌ترتیب افزایش در مقدار گاما در جدول ۳ آورده شده است.

جدول ۲- مقایسه تأثیر پارامترهای مختلف بر مقدار گاما.

Table 2. Comparison of various parameters effect on Gamma value.

شماره ترکیب Combination number	متغیر حذف‌شده Removed Parameter	گاما Gamma (Cm d^{-1})	شیب Gradient	خطای استاندارد Standard error	نسبت V V ratio
1	-	0.0020	0.0090	0.0001	0.3010
2	طول و عرض جغرافیایی Longitude and latitude	0.0008	0.0349	0.0003	0.1205
3	شن Sand	0.0021	0.0083	0.0003	0.3098
4	سیلت Silt	0.0021	0.0082	0.0002	0.3102
5	رس Clay	0.0020	0.0095	0.0001	0.2982
6	کربن آلی Organic carbon	0.0020	0.0090	0.0001	0.3010
7	مواد خنثی‌شونده TNV	0.0021	0.0078	0.0001	0.3133
8	هدایت الکتریکی EC	0.0021	0.101	0.0001	0.3048
9	اسیدیته pH	0.0020	0.0105	0.0001	0.3026
10	رطوبت اشباع Saturation moisture	0.0014	0.0200	0.0004	0.2022
11	جرم مخصوص حقیقی Particle density	0.0022	0.0088	0.0001	0.3305
12	جرم مخصوص ظاهری Bulk density	0.0026	0.0041	0.0001	0.3823

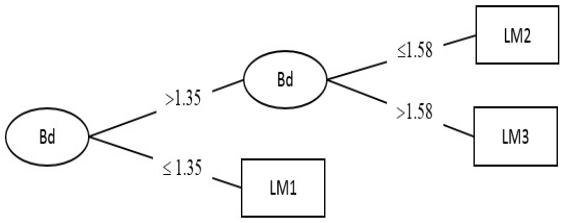
جدول ۳- مقدار گاما برای پنج ترکیب برتر.

Table 3. Gamma value of five most optimum combinations of input parameters.

شماره ترکیب Combination number	پارامترهای ترکیب Parameters of the combination	گاما Gamma
1	شن، سیلت، رس، مواد خنثی‌شونده، هدایت الکتریکی، جرم مخصوص ظاهری Sand, Silt, Clay, TNV, EC, Bulk density	0.00004215
2	شن، سیلت، رس، کربن آلی، مواد خنثی‌شونده، هدایت الکتریکی، جرم مخصوص ظاهری Sand, Silt, Clay, Organic carbon, TNV, EC, Bulk density	0.00004298
3	شن، سیلت، مواد خنثی‌شونده، هدایت الکتریکی، جرم مخصوص ظاهری Sand, Silt, TNV, EC, Bulk density	0.000118
4	شن، سیلت، کربن آلی، مواد خنثی‌شونده، هدایت الکتریکی، جرم مخصوص ظاهری Sand, Silt, Organic carbon, TNV, EC, Bulk density	0.000119
5	شن، رس، کربن آلی، مواد خنثی‌شونده، هدایت الکتریکی، جرم مخصوص ظاهری Sand, Clay, Organic carbon, TNV, EC, Bulk density	0.000311

براساس نتایج آزمون گاما و مقایسه مقادیر گاما برای ترکیبات به وجود آمده می‌توان ترکیب بهینه را معین کرد. با توجه به جدول ۳ ترکیب بهینه ترکیب شماره یک است که شامل پارامترهای شن، سیلت، رس، مواد خنثی‌شونده، هدایت الکتریکی و جرم مخصوص ظاهری خاک است و کم‌ترین مقدار گاما را در بین تمامی ترکیبات ممکنه دارد.

نتایج مدل‌سازی با درخت تصمیم: برای آموزش درخت تصمیم تعداد ۱۱۳ نمونه در نظر گرفته شد. الگوریتم مورد استفاده M5P بود و فرآیندهای هموارسازی و هرس کردن درخت برای مدل‌سازی استفاده شدند. ترکیب ورودی شامل درصد شن، درصد سیلت، درصد رس، درصد مواد خنثی‌شونده، هدایت الکتریکی و جرم مخصوص ظاهری (ترکیب بهینه از آزمون گاما) بود. پس از اجرای الگوریتم موردنظر نتایج به‌دست آمده در شکل ۱ و جدول ٤ آمده است. ساختار درختی ساده‌ای تشکیل شد که تنها متغیر جرم مخصوص ظاهری خاک را به‌عنوان مهم‌ترین متغیر دسته‌بندی‌کننده انتخاب کرده است. و سه مدل خطی در برگ‌های درخت ظاهر شد.



شکل ۱- ساختار درختی ایجاد شده توسط درخت تصمیم (Bd= جرم مخصوص ظاهری خاک بر حسب g/cm^3).

Figure 1. The tree structure made by Decision tree. (Bd= Bulk density of soil in g/cm^3).

جدول ٤- مدل‌های ساخته شده درخت تصمیم.

Table 4. Linear models constructed by Decision Tree.

مدل Model	هدایت هیدرولیکی اشباع Saturated hydraulic conductivity (Cm d^{-1})
LM1	$K_s = 0.5154* Silt – 0.2572* Clay + 0.243* TNV- 0.4693* Bd + 0.2254$
LM2	$K_s = 0.2234* Silt + 0.0528* Clay + 0.0922* Bd + 0.1231$
LM3	$K_s = 0.1189* Silt + 0.0713* Clay + 0.0343 * TNV – 0.1481* Bd + 0.1289$

(K_s= هدایت هیدرولیکی اشباع، Silt= سیلت، Clay= رس، TNV= درصد مواد خنثی‌شونده، Bd= جرم مخصوص ظاهری)

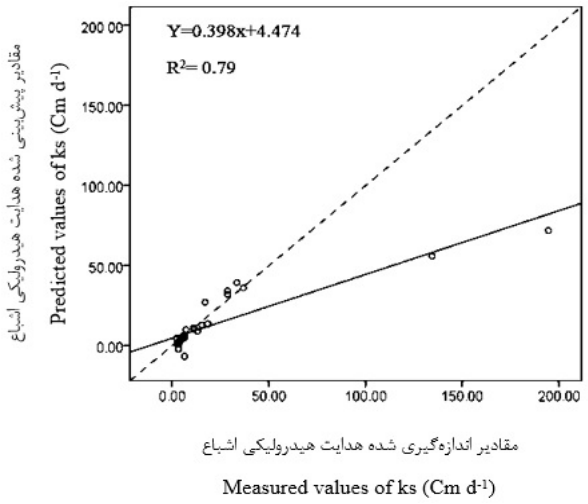
برای ارزیابی عملکرد مدل درخت تصمیم در برابر داده‌های جدید، با ۲۵٪ داده‌ها که برای آزمون مدل در نظر گرفته شده بود، مدل اجرا شد. معیارهای ارزیابی عملکرد مدل در هر دو مرحله شامل RMSE، MAE، R^2 و MAPE در جدول ۵ آورده شده است. مقادیر RMSE، MAE و R^2 آموزش مدل نشان‌دهنده عملکرد بسیار خوب مدل بود. در مرحله آزمون، دقت مدل برای پیش‌بینی داده‌های جدید به خوبی مرحله آموزش نیست و عملکردش کاهش یافته است.

جدول ۵- معیارهای ارزیابی درخت تصمیم در مرحله آموزش و آزمون.

Table 5. Evaluation statistics for Decision tree in training and testing steps.

MAPE	MAE	RMSE	R^2	
1.64	1.28	1.88	0.95	آموزش Training
20.50	7.42	23.89	0.79	آزمون Testing

ضریب تعیین مهم‌ترین معیاری است که به کمک آن می‌توان رابطه بین دو متغیر را توضیح داد و مقداری بین صفر و یک دارد ($0 \le R^2 \le 1$). شاخص ضریب تعیین (R^2) هرچه به یک نزدیک‌تر باشد بهتر است. در مرحله آزمون مدل، این مقدار برابر ۰/۷۹ به‌دست آمد که خوب است. برای دستیابی به اطلاعات بیش‌تر، نمودار پراکنش نقاط و بهترین خط برازنده به این نقاط رسم شد. این نمودار در شکل ۲ آمده است. با توجه به نمودار پراکنش ارتباط مستقیم خطی بین داده‌های اندازه‌گیری‌شده و پیش‌بینی‌شده مشاهده می‌شود و بیش‌تر نقاط حول خط رگرسیونی ۱:۱ (خط y=x) و نزدیک به آن هستند. نقاطی که زیر این خط مرجع قرار گرفته‌اند نشان‌دهنده این هستند که مدل مقدار آن‌ها را کم‌تر از مقدار واقعی تخمین زده است و نقاط بالاتر از این خط نشان می‌دهند که مدل مقدار آن‌ها را بیش‌تر تخمین زده است. معیار R^2 به تنهایی برای برآورد دقت مدل نمی‌تواند مناسب باشد، زیرا ممکن است مدلی با وجود R^2 بالا، مقادیر واقعی را اشتباه پیش‌بینی کرده باشد.

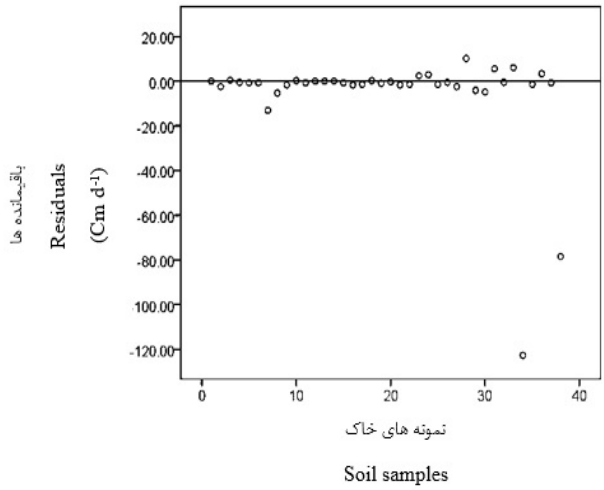


شکل ۲- پراکنش داده‌های اندازه‌گیری‌شده و پیش‌بینی‌شده مدل درخت تصمیم.

Figure 2. Measured and predicted values scatter for the Decision Tree model.

شاخص RMSE یا جذر میانگین مربعات خطا و MAE یا میانگین قدرمطلق خطا دو معیار ارزیابی کیفیت مدل براساس بررسی میزان خطای پیش‌بینی هستند. مقدار به‌دست آمده برای RMSE برابر ۲۳/۸۹ سانتی‌متر بر روز است. به‌طورکلی مقدار RMSE از صفر تا بینهایت می‌تواند باشد اما هرچه مقدار

RMSE به صفر نزدیک‌تر باشد بهتر است. MAE برابر ۷/٤۲ سانتی‌متر بر روز به‌دست آمد. تفاوت این دو معیار در این است که RMSE نسبت به MAE، برای خطاهای بزرگ‌تر جریمه سنگین‌تری قائل است (۳). نمودار پراکنش باقی‌مانده‌ها ($k_{si} - k_{oi}$) برای مدل در شکل ۳ آمده است. با توجه به این نمودار مشخص می‌شود که باقی‌مانده‌ها اکثراً نزدیک به صفر هستند و پیش‌بینی دو نقطه با خطای زیادی مواجه بوده است. خطای ناشی از پیش‌بینی این دو نمونه به‌ویژه منجر به افزایش زیادی در RMSE شده است. مقدار MAPE برابر ۲۰/۵۰ درصد به‌دست آمده که این معیار نیز نشان می‌دهد مدل دقت بالایی در برآورد هدایت هیدرولیکی اشباع ندارد و پیش‌بینی آن با خطای بالایی مواجه بوده.



شکل ۳- پراکنش باقی‌مانده‌ها برای مدل درخت تصمیم.

Figure 3. Residuals of the Decision Tree model.

نتایج مدل‌سازی با روش IBk: با تعداد ۱۱۳ نمونه یا سری داده‌های آموزشی (مرجع)، مقدار مناسب k از طریق ارزیابی تقاطعی یکی بیرون پیدا شد. به این صورت که ریشه دوم تعداد نمونه‌ها محاسبه شد و به‌عنوان بیش‌ترین مقدار k در نظر گرفته شد (رابطه ۱۰).

$$k_{\max} = \sqrt{113} = 10.63 \qquad (۱۰)$$

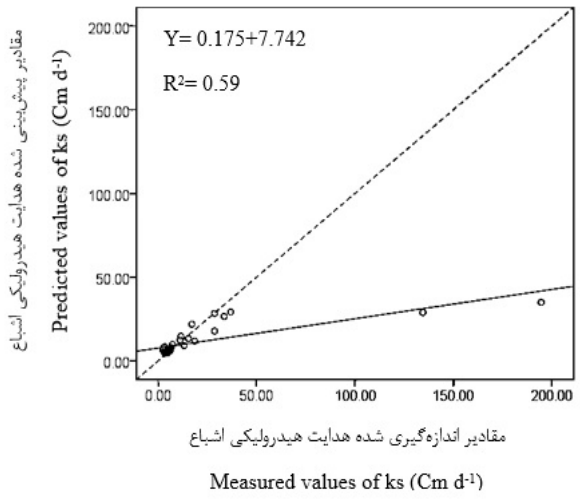
به‌ازای هر مقدار k بین دو عدد یک و ۱۱، میانگین مربعات خطا توسط خود الگوریتم محاسبه شد و مقدار مناسب برای k برابر ٤ به‌دست آمد. دو نوع تابع وزن‌دهی استفاده شد. معیارهای ارزیابی مدل در این مرحله به‌دست آمد و نتایج در جدول ٦ آورده شده است. مرحله آموزش و آزمون IBk کمی متفاوت از سایر روش‌ها مانند درخت تصمیم است. آزمون مدل با تعداد ۳۸ نمونه صورت گرفت. نمونه‌های این سری آزمایشی به‌عنوان نمونه هدف قرار گرفتند و ٤ تا از نزدیک‌ترین همسایه‌ها از نمونه‌های آموزشی که در مرحله قبل وارد مدل شده بود پیدا شدند. برای هر یک از نمونه‌های این سری نیز تخمین هدایت هیدرولیکی اشباع صورت گرفت. معیارهای ارزیابی هر یک از مدل‌های IBk، IBk با وزن‌دهی معکوس فاصله و IBk با وزن‌دهی خطی در جدول ٦ آورده شده است.

جدول ٦- مقایسه معیارهای ارزیابی IBk و دو روش متفاوت وزن‌دهی فاصله در مرحله آموزش و آزمون.

Table 6. Comparison of evaluation statistics for IBk and two distance weighting methods in training and testing steps.

آزمون Testing			آموزش Training			
بدون وزن‌دهی فاصله No distance weighting	$1/d_i$	$1-d_i$	بدون وزن‌دهی فاصله No distance weighting	$1/d_i$	$1-d_i$	
0.59	0.59	0.59	0.90	1	0.91	R^2
31.24	31.23	31.24	2.86	0.17	2.76	RMSE
9.44	9.32	9.45	1.88	0.11	1.83	MAE
24.14	23.24	24.13	7.37	0.18	6.99	MAPE

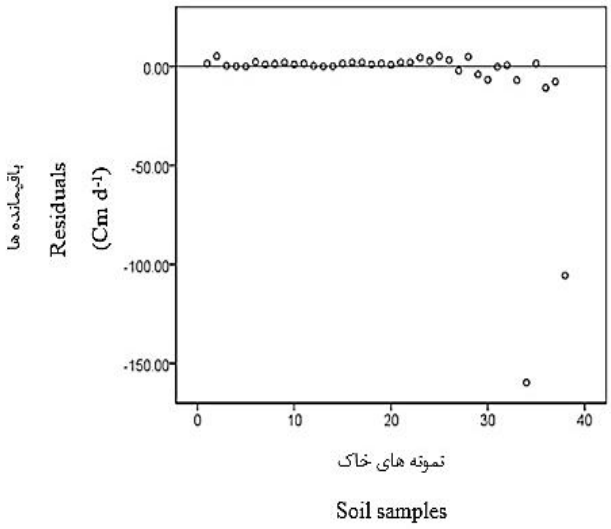
در مرحله آموزش، وزن‌دهی معکوس فاصله بیش‌ترین بهبود را در نتایج ایجاد کرد. پس از آن بهترین نتایج مربوط به نوع $1-d_i$ بود. اما در مرحله آزمون، نتایج به‌ویژه با توجه به معیار R^2 بهبود نیافت. در این مرحله نیز تابع معکوس فاصله، به مقدار بسیار کمی نتایج بهتری نسبت به دو مدل دیگر نشان داد و خطای RMSE، MAE و MAPE کم‌تری داشت. بنابراین می‌توان گفت مدل IBk به انتخاب نوع سیستم وزن‌دهی حساس نیست. نمس و همکاران (٢٠٠٦) نیز از سیستم وزن‌دهی متفاوتی با این پژوهش استفاده کردند و نشان دادند که k-NN به انتخاب نوع وزن‌دهی حساس نیست (٢٠). نتایج مدل‌سازی مرحله آموزش برای مدل IBk با وزن‌دهی معکوس، نشان از عملکرد بسیار خوب مدل دارد. معیار R^2 به‌دست آمده از مدل در مرحله آزمون برابر ٠/٥٩ است. نمودار پراکنش نقاط مقادیر مشاهده‌شده و پیش‌بینی‌شده و معادله خط رگرسیونی که به این نقاط برازش داده شده در شکل ٤ آورده شد. بیش‌تر نقاط نزدیک خط مرجع y=x قرار گرفته‌اند. مقدار RMSE برابر با ٣١/٢٣ سانتی‌متر بر روز به‌دست آمده و نشان می‌دهد که دقت این مدل پایین بوده است.



شکل ٤- نمودار پراکنش داده‌های اندازه‌گیری‌شده در برابر پیش‌بینی‌شده مدل IBk.

Figure 4. Measured and predicted values scatter plot for IBk.

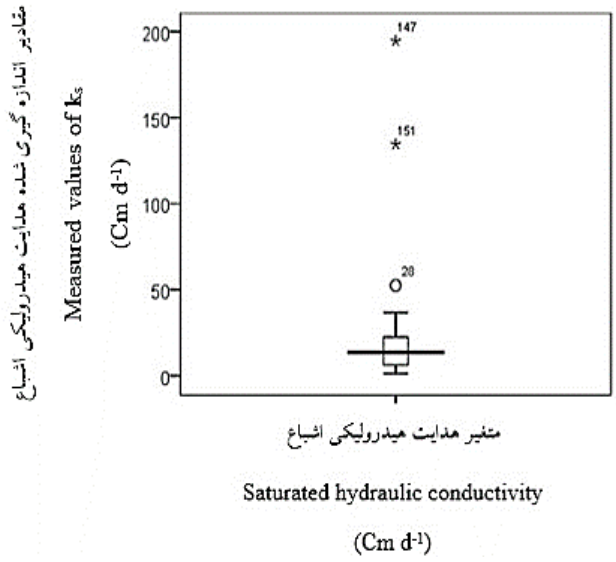
با توجه به نمودار پراکنش باقی‌مانده‌ها که در شکل ۵ آمده است، مشخص می‌شود که همانند مدل درخت تصمیم، پیش‌بینی دو نمونه با خطای زیادی مواجه بوده است. بالاتر بودن مقدار RMSE می‌تواند ناشی از پیش‌بینی این داده‌ها باشد.



شکل ۵- پراکنش باقی‌مانده‌ها برای مدل IBk.

Figure 5. Residuals of IBK model.

با توجه به نمودار پراکنش باقی‌مانده‌ها برای هر دو مدل M5P و IBk (شکل‌های ۳ و ۵) دیده می‌شود که پیش‌بینی دو نقطه با خطای بسیار زیادی مواجه بوده است و باعث دقت کم‌تر مدل‌ها در پیش‌بینی داده‌های آزمون شده است. نمودار جعبه‌ای توزیع هدایت هیدرولیکی اشباع نمونه‌های خاک برای کمک به تشخیص داده‌های پرت تک‌متغیره رسم شد که در شکل ٦ آمده است. سه نقطه‌ای که با علامت ستاره و دایره مشخص شده‌اند نشان‌دهنده نقاط یا داده‌های پرت هستند (اعداد درون نمودار نشان‌دهنده شماره نمونه هستند). دو نمونه ١٤٧ و ١٥١ که بیش‌ترین هدایت هیدرولیکی اشباع را به‌ترتیب برابر ١٣٤/٦ و ١٩٤/٦٢ سانتی‌متر بر روز داشتند در آزمون مدل استفاده شده و نمونه ٢٨ با مقدار k_s برابر با ٥٢/٥١ در آموزش مدل استفاده شده است.

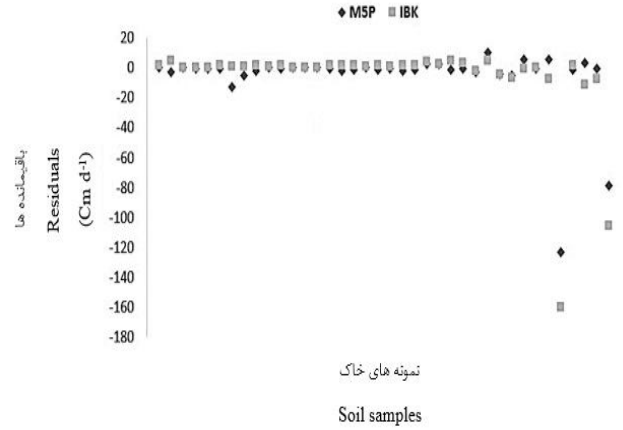


شکل ٦- توزیع هدایت هیدرولیکی اشباع نمونه‌های خاک.

Figure 6. Distribution of saturated hydraulic conductivity of soil samples.

داده پرت، نمونه‌ای است که با مقادیر سایر نمونه‌ها در یک بانک داده متفاوت باشد. نمونه‌های پرت به دلایل متعددی مانند خطای اندازه‌گیری، ثبت یا وارد کردن داده‌ها می‌توانند ایجاد شده باشند یا در صورت صحیح بودن اندازه‌گیری، از جامعه‌ای متفاوت هستند و بیانگر یک اتفاق نادر می‌باشند (٤). خامیس و همکاران (۲۰۰۵) نشان داده‌اند که داده‌های پرت سبب کاهش عملکرد شبکه‌های عصبی مصنوعی هم برای آموزش و هم برای آزمون مدل شده است (۱۳). در مواجه با داده پرت باید دید که این داده‌ها واقعی هستند یا اشتباهند به‌عبارتی در طبیعت رخ می‌دهند یا نه. هدایت هیدرولیکی اشباع دامنه گسترده‌ای از نظر تغییرات دارد و مقدار آن بسته به نوع خاک از چند میلی‌متر در روز تا چند متر در روز متغیر است. مقدار k_s برای هر سه نمونه ذکر شده در نمودار شکل ٦ مقادیری طبیعی و در همین طیف هستند اما با سایر نمونه‌ها تفاوت زیادی دارند. به‌دلیل گستردگی مقادیر هدایت هیدرولیکی اشباع و تغییرات زیاد آن، این نمونه‌ها حذف نشدند.

مقایسه دو مدل: برای مقایسه IBk با درخت تصمیم، مدل IBk با وزن‌دهی معکوس فاصله استفاده شد. با توجه به معیارهای R^2، RMSE، MAE و MAPE مدل درخت تصمیم دقت بالاتر و همبستگی بهتری در پیش‌بینی هدایت هیدرولیکی اشباع داشت. نمودار پراکنش باقی‌مانده‌ها برای هر دو مدل در شکل ۷ آمده است. هر دو مدل خطاهای نزدیک به صفر دارند و تقریباً مشابه هستند. در مدل درخت تصمیم پراکنش باقی‌مانده‌ها کم‌تر از IBk بوده و به صفر نزدیک‌ترند. هر دو مدل به داده‌های پرت حساس بوده اما درخت تصمیم برای پیش‌بینی داده‌های پرت باقی‌مانده کم‌تری داشته است. از آن‌جایی‌که هدایت هیدرولیکی اشباع یک ویژگی با تغییرات بالا است، از این‌رو درخت تصمیم مدل بهتری در برآورد هدایت هیدرولیکی اشباع است. با توجه به معیار MAPE، این تفاوت اندک است.



شکل ۷- مقایسه پراکنش باقی‌مانده‌ها در دو مدل درخت تصمیم (M5P) و IBk.

Figure 7. Comparison of residuals of the two models Decision Tree (M5P) and IBk.

نتیجه‌گیری

در این پژوهش از آزمون گاما برای انتخاب مهم‌ترین پارامترهای ورودی و ترکیبی بهینه از پارامترها برای مدل‌سازی هدایت هیدرولیکی اشباع به روش درخت تصمیم و یادگیری برپایه نمونه استفاده شد. نتایج آزمون گاما نشان داد که علاوه بر پارامترهای توزیع اندازه ذرات خاک و وزن مخصوص ظاهری که در بیش‌تر پژوهش‌های پیشین به‌عنوان متغیرهای ورودی مدل استفاده شدند، پارامترهای هدایت الکتریکی و درصد مواد خنثی‌شونده نیز می‌توانند برای پیش‌بینی هدایت هیدرولیکی اشباع مورد استفاده واقع گردند. نتایج به‌دست آمده از دو مدل درخت تصمیم و یادگیری برپایه نمونه نشان دادند که این دو مدل دقت پیش‌بینی بالایی نداشتند. وجود داده‌های پرت در مرحله آزمون بر عملکرد پایین مدل‌ها مؤثر بوده. داده‌های آموزشی که برای آموزش مدل‌ها استفاده شد از نظر خصوصیات زودیافت خاک و هدایت هیدرولیکی اشباع نمونه‌ها دامنه محدودی از مقادیر را داشت. بنابراین برای دست‌یابی به دقت پیش‌بینی بالاتر در مدل‌سازی لازم است که از نمونه‌هایی استفاده شود که دارای تنوع بالا در بافت خاک باشند و از نظر خصوصیات زودیافت و هدایت هیدرولیکی اشباع، دامنه گسترده‌تری از مقادیر را شامل شوند.

منابع

1.Abbasi, F. 2017. Advanced soil physics. Tehran university press, 320p. (In Persian)
2.Aha, D.W., Kibler, D., and Albert, M.K. 1991. Instance-based learning algorithms. Machine learning, 6: 37-66.
3.Azar, A., and Momeny, M. 2006. Statistics and its application in management (Statistical analysis). Tehran: The organization for researching and composing university textbooks in the Humanities (SAMT). 440p. (In Persian)
4.Cateni, S., Colla, V., and Vannucci, M. 2008. Outlier detection methods for industrial applications. In: Arámburo, A. and Ramírez Treviño, A. (eds), Advances in Robotics, Automation and Control. (265-282). In Tech, Vienna, Austria.
5.Debeljak, M., and Džeroski, S. 2011. Decision Trees in Ecological Modelling. In: Jopp, F., Reuter, H., Breckling, B. (eds), Modelling Complex Ecological Dynamics. (197-209). Springer, Berlin, Heidelberg.

6.Evans, D. 2002. The Gamma Test: Data-derived estimates of noise for unknown smooth models using near-neighbour asymptotics. Doctoral thesis, Department of computer science, Cardiff university, University of Wales.
7.Ghabaei Sough, M., Masaedi, A., Hesam, M., and Hezarjaribi, A. 2010. Evaluation effect of input parameters preprocessing in Artificial Neural Networks (Anns) by using stepwise regression and Gamma test techniques for fast estimation of daily evapotranspiration. J. Water Soil. 24: 3. 610-624. (In Persian)
8.Haghverdi, A., Ghahraman, B., Khoshnood Yazdi, A.A., and Arabi, Z. 2010. Estimating of water content in FC and PWP in North and North East of Iran's soil samples using k-Nearest Neighbor and Artificial Neural Networks. J. Water Soil. 24: 4. 804-814. (In Persian)
9.Jabro, J.D. 1992. Estimation of saturated hydraulic conductivity of soils from particle size distribution and bulk density data. Transactions of the ASAE, 35: 2. 557-560.
10.Jalali, V.R., and Homaee, M. 2011. Introducing a nonparametric model using k-nearest neighbor technique for predicting soil bulk density. Journal of Science and Technology of Agriculture and Natural Resources, Water and Soil Science. 15: 56. 181-191. (In Persian)
11.Jones, A.J. 1998. The WinGamma user guide. University of Wales, Cardiff.
12.Kemp, S.E., Wilson, I.D., and Ware, J.A. 2005. A tutorial on the gamma test. J. Sim. Syst. Sci. Technol. 6: 1-2. 67-73.
13.Khamis, A., Ismail, Z., Haron, Kh., and Tarmizi Mohammad, A. 2005. The effects of outlier data on neural network performance. J. Appl. Sci. 5: 8. 1394-1398.
14.Khashei Siuki, A., Jalali Moakhar, V.R., Noferesti, A.M., and Ramazani, Y. 2015. Comparing nonparametric k-nearest neighbor technique with ANN model for predicting soil saturated hydraulic conductivity. J. Soil Manage. Sust. Prod. 5: 3. 81-95. (In Persian)
15.Lall, U., and Sharma, A. 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research, 32: 3. 679-693.
16.Mahdian, M.H. 2005. Soil hydraulic conductivity and its application in drainage designs. J. Agric. Engin. Res. 6: 23. 159-170. (In Persian)
17.Mallant, D., Mohanty, B.P., Vervoort, A., and Feyen, J. 1997. Spatial analysis of saturated hydraulic conductivity in a soil with macropores. Soil Technology. 10: 115-131.
18.Moghaddamnia, A., Gousheh, M.G., Piri, J., Amin, S., and Han, D. 2009. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Advances in Water Resources. 32: 1. 88-97.
19.Moncada, M.P., Gabriels, D., and Cornelis, W.M. 2014. Data-driven analysis of soil quality indicators using limited data. Geoderma. 235: 271-278.
20.Nemes, A., Rawls, W.J., and Pachepsky, Y.A. 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Amer. J. 70: 2. 327-336.
21.Nosrati Karizak, F., Movahedi Naeni, S.A., and Hezarjaribi, A. 2012. Using Artificial Neural Networks to estimate saturated hydraulic conductivity from easily available soil properties. J. Soil Manage. Sust. Prod. 2: 1. 95-110. (In Persian)
22.Rasoulzadeh, A., Razavi, S., and Neyshoubori, R. 2012. Evaluation the accuracy of methods of estimating saturated hydraulic conductivity in different soils. J. Water Res. Agric. 26: 3. 303-316. (In Persian)
23.Schaap, M.G., Leij, F.J., and Van Genuchten, M.T. 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251: 3-4. 163-176.
24.Torabi, M. 2004. Assessment of five methods of saturated hydraulic conductivity measurement in a saline soil. 2nd Students Conference on Soil and Water Resources. University of Shiraz. (In Persian)
25.Wang, Y., and Witten, I.H. 1997. Inducing model trees for continuous classes. In Proceedings of the Ninth European Conference on Machine Learning. Pp: 128-137.

Gorgan University of Agricultural Sciences and Natural Resources

J. of Water and Soil Conservation, Vol. 25(5), 2019
http://jwsc.gau.ac.ir
DOI: 10.22069/jwsc.2018.12334.2691

Comparing decision tree and instance-based learning models to estimate soil saturated hydraulic conductivity

M. Farzadmehr[1], *M. Dastourani[2] and A. Khashei-Siuki[3]
[1]M.Sc. Graduate, Dept. of Water Science and Engineering, University of Birjand,
[2]Assistant Prof., Dept. of Water Science and Engineering, University of Birjand,
[3]Associate Prof., Dept. of Water Science and Engineering, University of Birjand
Received: 05.29.2018; Accepted: 09.18.2018

Abstract

Background and Objectives: Soil saturated hydraulic conductivity is one of the most important physical characteristics of soils which affects water movement in soil. Knowledge of this parameter can help to understand and solve environmental problems. However, measurement of this parameter by direct laboratory and field methods is hard, time consuming and expensive. Thus, there is need to use alternative methods based on conveniently available soil properties to estimate it with less effort, time and cost. Nonparametric methods are new indirect methods to estimate hydraulic properties of soil, including soil saturated hydraulic conductivity (k_s). The aim of this study was to use two methods such as M5P decision tree and an IBk instance-based learning method, which is a classifier with k nearest neighbors to estimate k_s from conveniently available properties of soil.

Materials and Methods: In this study a dataset of 151 soil samples which was collected from a site in Bojnord province was used. Conveniently available soil properties includig sand, silt and clay percentage, bulk density, particle density, EC, OC, TNV, saturated moisture and pH. Saturated hydraulic conductivity was measured with the Guelph permeameter. The Gamma test was used to determine important parameters for predicting and the modeling procedure of k_s. Then, various combinations of parameters of the data set were compared to each other based on their Gamma value, to determine the optimum combination of parameters for modeling k_s. Using the optimum combination which had the least Gamma value, the M5P decision tree and the IBk instance-based learning methods were performed. To improve the IBk, two different distance weighting systems were used. Finally, evaluation statistics of each model including R^2, RMSE, MAE and MAPE were calculated.

Results: The optimum combination determined by the Gamma test which was then used for modeling, included sand, silt and clay percent, TNV percent, EC and bulk density. The tree selected bulk density as the most important discriminative parameter and constructed 3 linear equations for predicting k_s, based on the bulk density value. Evaluation criteria calculated for this model with RMSE= 23.89 cm/d and MAPE= 20.50% didn't predict k_s accurately. Different weighting systems didn't improve IBk performance. Also, the IBk model with RMSE=31.23 cm/d and MAPE=23.24% didn't estimate k_s accurately.

Conclusion: The decision tree model performed better than the instance-based learning model to estimate k_s. Also, the tree showed some information about the structure of the studied soil.

***Keywords**:* Gamma test, IBk algorithm, M5P algorithm

* Corresponding Author; Email: mdastourani@birjand.ac.ir