Iranian Journal of
Rainwater Catchment Systems
مجله علمی سامانه‌های سطوح آبگیر باران
شاپا الکترونیکی: ۱۵۳۱-۲۷۸۲ شاپا چاپی: ۵۹۷۰-۲۴۲۲

پیش‌بینی بارش روزانه ایستگاه سردشت با استفاده از الگوریتم‌های تنبل و مدل‌های درختی

میلاد شرفی*[۱]، جواد بهمنش[۲]

۱. دانشجوی دکتری، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه ارومیه، ارومیه، ایران.
۲. استاد گروه مهندسی آب، دانشکده کشاورزی، دانشگاه ارومیه، ارومیه، ایران.

تاریخ دریافت: ۱۴۰۱/۰۷
تاریخ پذیرش: ۱۴۰۱/۰۹
صفحات: ۱۰-۱

نوع مقاله: علمی-پژوهشی

چکیده

با توجه به توزیع ناهمگون بارش، پیش‌بینی وقوع آن‌یکی از راه‌کارهای اولیه و اساسی برای پیش‌گیری از بلایای احتمالی و خسارات ناشی از آن است. با توجه به بالا بودن میزان بارندگی در شهرستان سردشت، روی آوردن مردم این شهرستان به کشاورزی در سال‌های اخیر و عدم استفاده از مدل‌های طبقه‌بندی در ایستگاه مورد مطالعه، پیش‌بینی هرچه دقیق‌تر پارامتر بارش روزانه امری ضروری است. از طرفی دیگر، با این‌که عملکرد مطلوب الگوریتم‌های تنبل و مدل‌های درختی باعث افزایش استفاده از آن‌ها برای پیش‌بینی پدیده‌های مختلف هیدرولوژیکی شده اما این الگوریتم‌ها در شهرستان سردشت مورد استفاده قرار نگرفته‌اند. لذا در این پژوهش، چهار مدل Kstar، M5P، الگوریتم یادگیری با وزن‌دهی محلی و جنگل تصادفی برای پیش‌بینی بارش روزانه ایستگاه سردشت به کار گرفته‌شده است. در این مطالعه از هفت پارامتر ورودی میانگین دما، حداکثر دما، رطوبت نسبی متوسط، حداکثر رطوبت نسبی، سرعت باد متوسط، حداکثر سرعت باد و ساعات آفتابی که هم‌زمان با بارش روزانه بودند، برای مدل‌ها استفاده شد. مقایسه و ارزیابی بین پارامترهای ورودی نشان داد که پارامتر ساعات آفتابی ازجمله مهم‌ترین پارامترهای ورودی بوده که نقش قابل‌توجهی در دقت پیش‌بینی مدل‌های مورد استفاده داشته است. نتایج به‌دست‌آمده نشان داد که مدل درختی M5P در سناریوی هفتم بهترین عملکرد را با بیش‌ترین ضریب همبستگی (۰/۷۳۴ میلی‌متر بر روز) نسبت به دیگر مدل‌ها داشته است. هم‌چنین، سناریوی هفتم عملکرد بالایی نسبت به بقیه سناریوها از خود نشان داد. لذا می‌توان گفت که افزایش ورودی مدل‌ها تا حدودی رابطه مستقیمی با دقت آن‌ها دارد. به‌طورکلی می‌توان گفت که مدل درختی M5P برای مدل‌سازی و پیش‌بینی بارش روزانه شهرستان سردشت مناسب بوده و برای استفاده‌های بعدی پیشنهاد می‌شود.

کلمات کلیدی: الگوریتم یادگیری، پیش‌بینی، سردشت، مدل‌سازی، مدل درختی.

مقدمه

بارش نقش مهمی در چرخه آب و انرژی جهانی دارد. بارش فعالیت‌های کشاورزی، اکولوژی و محیط‌زیست یک منطقه را تعیین می‌کند. ازاین‌رو به‌عنوان عامل کلیدی توسعه اجتماعی و اقتصادی هر منطقه محسوب می‌شود (Pour et al., 2020). درعین‌حال، بارش تأثیرگذارترین عامل برای انواع مخاطرات طبیعی است (Wright et al., 2017). بارش بیش‌ازحد و سیل شایع‌ترین خطرات طبیعی در سراسر زمین هستند (Belachsen et al., 2017). کمبود بارش و خشک‌سالی ویرانگرترین بلایا ازنظر خسارت‌های اقتصادی هستند (Brito et al., 2018). هم‌چنین، بارش عامل تعیین‌کننده بسیاری از انواع دیگر مخاطرات طبیعی مانند رانش زمین، فرسایش خاک و فرونشست رودخانه است (Pourghasemi et al., 2020). بنابراین، پیش‌بینی بارش موضوع اصلی موردتوجه هیدرولوژیست‌ها و دانشمندان برای چندین دهه است (Praveen et al., 2020). بااین‌حال بیش از ۵۴ درصد جمعیت جهان در مناطقی زندگی می‌کنند که بحران آب جدی است (Bostan et al., 2012). ایران در یک کمربند خشک قرار دارد و میانگین بارش آن تنها معادل یک‌پنجم میانگین جهانی است (Bozorg-Haddad et al., 2020; Deh-Haghi et al., 2020; Dehghani et al., 2020). به‌طور طبیعی، شناخت عوامل مؤثر بر

[1] *Email: miladsharafi1@gmail.com نویسنده مسئول: میلاد شرفی

[DOR: 20.1001.1.24235970.1401.10.3.1.4]

الگوهای بارش در چنین مناطقی بسیار حیاتی است. باوجوداین، پیش‌بینی بارش دشوار است زیرا تغییرات مکانی و زمانی زیادی را به نمایش می‌گذارد. الگوهای ذاتی غیرخطی و پیچیده بارش، آن را به یک کاربرد جذاب برای شبیه‌سازی تبدیل کرده است (Asghari & Nasseri, 2015; Wang et al., 2021). در سال‌های اخیر، با توجه به توانایی بالای الگوریتم‌های تنبل KSTAR، الگوریتم یادگیری با وزن‌دهی محلی (LWL)[1]، مدل‌های درختی M5P و جنگل تصادفی (RF)[2]، برای پیش‌بینی پدیده‌های مختلف هیدرولوژیکی زیاد مورد استفاده قرارگرفته‌اند. برای نمونه، Sihag و همکاران (۲۰۲۱) به ارزیابی رگرسیون درختی در برآورد دبی حوضه رودخانه پرداختند. نتایج این پژوهش نشان داد مدل M5P با دقت بالایی (ضریب همبستگی ۰/۸۷) دبی حوضه رودخانه را برآورد کرده است. Di Nunno و همکاران (۲۰۲۲) در تحقیقی به پیش‌بینی بارش در شمال بنگلادش با استفاده از مدل یادگیری ماشین ترکیبی M5P-SVR پرداختند. نتایج این بررسی نشان داد که مدل M5P-SVR منجر به بهترین پیش‌بینی‌ها در بین مدل‌های مورد استفاده در این مطالعه با مقادیر ضریب تعیین[3] ۰/۸۷ و ۰/۹۲ به‌ترتیب برای ایستگاه‌های Rangpur و Sylhet شد. پورصالحی و همکاران (۱۴۰۰) در مطالعه‌ای به بررسی عملکرد الگوریتم جنگل تصادفی در پیش‌بینی نوسانات سطح ایستابی آبخوان آزاد دشت بیرجند در مقایسه با دو مدل درخت تصمیم و شبکهٔ عصبی مصنوعی پرداختند. نتایج این مطالعه نشان داد شبیه‌سازی با استفاده از الگوریتم جنگل تصادفی براساس ضریب تعیین معادل ۰/۷۱۴ نشان داد که این الگوریتم توانایی نسبتاً زیادی در شبیه‌سازی تراز سطح ایستابی آبخوان دارد. Adnan و همکاران (۲۰۲۱) در تحقیقی جریان در حوزه آبریز Jhelum در پاکستان را با استفاده از الگوریتم یادگیری محلی وزنی پیش‌بینی کردند. نتایج این تحقیق نشان داد که مدل یادگیری محلی وزنی با الگوریتم افزایشی (LWL-AR[4]) با داشتن ضریب همبستگی ۰/۸۶ به‌عنوان بهترین مدل شناخته شد. He و همکاران (۲۰۲۲) در مطالعه‌ای به پیش‌بینی آلودگی نیترات آب‌های زیرزمینی و ارزیابی عوامل تأثیرگذار اصلی آن در منطقه Yinchuan واقع در شمال غربی چین با استفاده از جنگل تصادفی پرداختند. نتایج این بررسی نشان داد که پس از کالیبراسیون، مدل جنگل تصادفی در مرحله اعتبارسنجی با داشتن دقت ۰/۹۶۱ دارای دقت بالایی در پیش‌بینی آلودگی نیترات بوده است. در مطالعه‌ای دیگر، Bushara (۲۰۱۹) به پیش‌بینی آب‌وهوا با استفاده از مدل‌های محاسباتی KSTAR و M5P پرداخت که نتایج نشان‌دهنده عملکرد بالای مدل‌ها در پیش‌بینی بوده است. در مطالعه‌ای دیگر توانایی روش جنگل تصادفی (RF) در پیش‌بینی جریان جاری در زمان سررسید یک‌روزه در ۸۶ حوضه در شمال غربی اقیانوس آرام بررسی شد که نتایج حاکی از عملکرد بالای جنگل تصادفی (RF) در پیش‌بینی جریان بود (Pham et al., 2021). در مطالعه‌ای دیگر محققان به پیش‌بینی سطح آب روزانه دریاچه زربر با مقایسه بین مدل‌های M5P، جنگل تصادفی (RF) پرداختند که نتایج مطالعه نشان داد این مدل‌ها ابزارهای مقرون‌به‌صرفه‌ای برای پیش‌بینی‌های آینده هستند. هم‌چنین، Khosravi و همکاران (۲۰۲۰) به مدل‌سازی تصادفی آلودگی فلوراید آب‌های زیرزمینی با الگوریتم‌های تنبل پرداختند. نتایج این مطالعه نشان داد که الگوریتم‌ها قادر بودند مدل‌سازی را با دقت بالایی انجام دهند. با توجه به توسعه کشاورزی، میزان بالای بارش و آسیب‌های ناشی از سیلاب در سال‌های اخیر در ایستگاه سردشت پیش‌بینی هرچه دقیق‌تر بارش کمک شایانی به محققان در اخذ تصمیم‌های مناسب خواهد کرد. این مطالعه برای اولین بار بارش روزانه ایستگاه سردشت را با استفاده از داده‌های هواشناسی و با به‌کارگیری الگوریتم‌های درختی و تنبل پیش‌بینی می‌کند. باوجود میزان بالای بارش و خسارات ناشی از سیل و تگرگ، این منطقه همچنان موردتوجه پژوهش‌گران نبوده و مطالعات انگشت‌شماری به نقش مهم پارامتر بارش پرداخته است. بنابراین، این پژوهش برای اولین بار به پیش‌بینی بارش روزانه در ایستگاه سردشت با استفاده از روش‌های Kstar، الگوریتم یادگیری با وزن‌دهی محلی (LWL)، M5P و جنگل تصادفی (RF) پرداخته‌شده است.

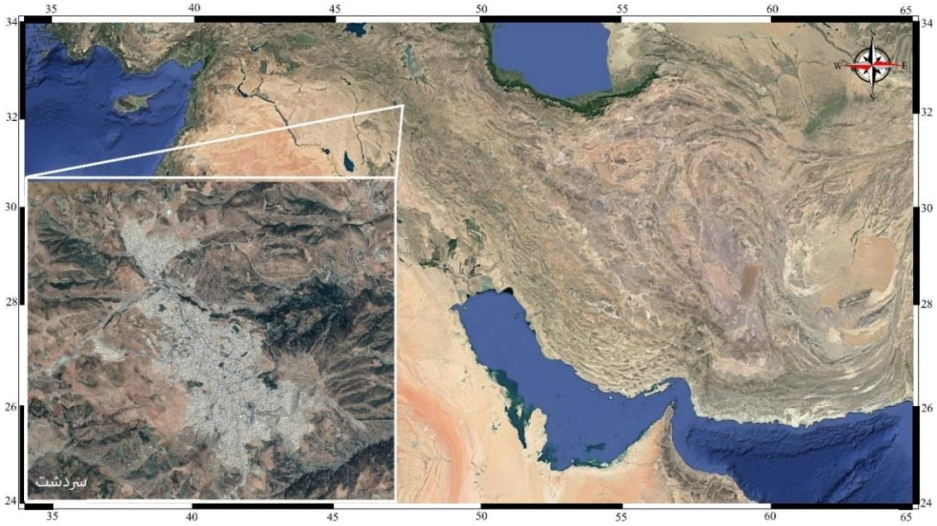
[1] Locally weighted learning
[2] Random forest
[3] Coefficient of determination
[4] Locally weighted learning- Additive Regression

[DOR: 20.1001.1.24235970.1401.10.3.1.4]

مواد و روش‌ها

منطقه مورد مطالعه

شهرستان سردشت با ارتفاع ۱۴۸۰ متر از سطح دریا و طول جغرافیایی ۴۵/۸۳ شرقی و عرض ۳۶/۲۵، در غرب کشور ایران و جنوب‌غربی استان آذربایجان غربی واقع‌شده است. میانگین بارش سالانه حدود ۷۹۰ میلی‌متر بوده و بیش‌ترین میزان بارش‌های استان آذربایجان غربی در این منطقه به وقوع می‌پیوندد (Ostad-Ali-Askari, 2020)، کشاورزی از اصلی‌ترین مشاغل ساکنان منطقه سردشت است.



شکل (۱): نقشه موقعیت جغرافیایی شهرستان سردشت

جدول (۱) سناریوهای مختلفی را که به‌عنوان ورودی و خروجی مدل‌ها در نظر گرفته‌شده‌اند، نشان می‌دهد. پارامترهای میانگین دما (T_{avg})، حداکثر دما (T_{max})، رطوبت نسبی متوسط (RH_{avg})، حداکثر رطوبت نسبی (RH_{max})، سرعت باد متوسط(U_{2ave})، حداکثر سرعت باد (U_{2max})، ساعات آفتابی (SSH) به‌عنوان پارامترهای ورودی و بارش روزانه (P) به‌عنوان پارامتر خروجی می‌باشند. نحوه ترکیب پارامترهای ورودی بر اساس ضریب همبستگی بوده، به‌طوری‌که از ضریب همبستگی کم‌تر به ضریب همبستگی بیش‌تر نوشته‌شده‌اند. پس از حذف داده‌های پرت از تمام داده‌های ۳۵ سال دوره آماری ۲۰۲۰-۱۹۸۵ در ایستگاه سردشت، از ۷۰ درصد داده‌ها برای آموزش مدل‌ها (۶۳۰۰ داده) و از ۳۰ درصد داده‌ها (۲۷۰۰ داده) برای مرحله آزمون استفاده شد. شکل (۱) نقشه موقعیت جغرافیایی شهرستان سردشت را نشان می‌دهد.

جدول (۱): ترکیب‌های مختلف ورودی، برای تخمین مقدار بارش روزانه در مدل‌های مورد مطالعه

سناریو	پارامترهای ورودی							خروجی
1	Tavg							P
2	Tavg	Tmax						P
3	Tavg	Tmax	RHavg					P
4	Tavg	Tmax	RHavg	RHmax				P
5	Tavg	Tmax	RHavg	RHmax	U2ave			P
6	Tavg	Tmax	RHavg	RHmax	U2ave	U2max		P
7	Tavg	Tmax	RHavg	RHmax	U2ave	U2max	SSH	P

الگوریتم kstar

الگوریتم kstar را می‌توان به‌عنوان روشی برای تجزیه‌وتحلیل خوشه‌ای تعریف کرد که عمدتاً هدف آن تقسیم "n مشاهداتی" به خوشه‌های k است که در آن‌ها هر مشاهده با نزدیک‌ترین میانگین به خوشه تعلق دارد. ما می‌توانیم الگوریتم kstar را به‌عنوان یک یادگیرنده مبتنی بر نمونه توصیف کنیم که از آنتروپی به‌عنوان اندازه‌گیری فاصله استفاده می‌کند. kstar یک طبقه‌بندی کننده ساده و مبتنی بر نمونه، شبیه به نزدیک‌ترین همسایگی (KNN)[1] است (Vijayarani

[1] K-Nearest Neighbour

& Muthulakshmi, 2013). یک روش طبقه‌بندی مبتنی بر مثال است که بر اساس موارد آموزشی مشابه، طبقه‌بندی را انجام می‌دهد و در مقایسه با الگوریتم‌های یادگیری ماشین، نتایج مطلوبی را حاصل می‌نماید. این روش برخلاف دیگر روش‌های داده‌کاوی که بر اساس تابع فاصله مبنی بر آنتروپی، طبقه‌بندی را انجام می‌دهند، از تابع شباهت برای تخمین متغیرهای مختلف استفاده می‌کند. فرض اصلی طبقه‌بندی مبتنی بر مثال، این است که موارد مشابه‌کلاس‌های مشابه داشته باشد (Cleary & Trigg, 1995). در این الگوریتم x، به کلاسی که بیش‌تر در میان نزدیک‌ترین نقاط رخ می‌دهد، اختصاص داده می‌شود، y_i که در آن $i = 1,2\ldots,k$. سپس از فاصله آنتروپیک برای بازیابی مشابه‌ترین نمونه‌ها استفاده می‌شود. با استفاده از فاصله آنتروپیک به‌عنوان یک متریک دارای تعدادی مزیت ازجمله مدیریت ویژگی‌های باارزش واقعی و مقادیر گم‌شده است (Vijayarani & Muthulakshmi, 2013). تابع KSTAR را می‌توان به‌صورت زیر محاسبه کرد:

$$K^*(y_i, x) = -\ln P^*(y_i, x) \quad (۱)$$

که P^* احتمال همه مسیرهای انتقالی از x تا y است.

الگوریتم یادگیری با وزن‌دهی محلی (LWL)

یک الگوریتم کلی برای یادگیری وزن‌دار شده به‌صورت محلی است. این الگوریتم با استفاده از روشی وزن‌ها را به نمونه‌ها نسبت می‌دهد و از روی نمونه‌های وزن‌دار شده، رده‌بندی را می‌سازد. این مدل برای مسائل طبقه‌بندی، رده‌بند Bayes Nave و برای مسائل رگرسیون، رگرسیون خطی انتخاب‌های خوبی هستند. می‌توان در این الگوریتم، تعداد همسایه‌های مورد استفاده را که پهنای باند هسته و شکل هسته مورد استفاده برای وزن‌دار کردن را (خطی، معکوس، یا گوسی) مشخص می‌کند، تعیین نمود (Joshuva & Sugumaran, 2020). LWL که تحت یک نوع یادگیری تنبل کار می کند، اغلب فرآیند آموزش را به تعویق می‌اندازد تا مقدار هدف یک مثال پرس‌وجو پیش‌بینی شود (Reyes et al., 2018).

الگوریتم M5P

مجموعه‌ای از نمونه‌های آموزشی (T) وجود دارد. هر نمونه آموزشی با مجموعه‌ای از ویژگی‌ها مشخص می‌شود که مقادیر ورودی هستند و دارای هدف متناظر که همان مقدار خروجی است. در این پژوهش، از الگوریتم M5P مشابه مطالعات موجود استفاده‌شده است (Quinlan, 1992; Solomatine & Dulal, 2003; Solomatine & SIEK, 2004). یک درخت تصمیم معمولاً از چهار بخش ریشه، شاخه، گره‌ها و برگ‌ها تشکیل‌شده است که گره‌ها با دایره نشان داده می‌شوند و شاخه‌ها نشان‌دهنده اتصال بین گره‌ها هستند. روش M5P قادر است مجموعه داده‌های بزرگ را به همراه داده‌های مفقودشده با تقسیم فضاهای ورودی به زیر فضاهای کوچک‌تر بازیابی کند. به‌طورکلی، حداقل تعداد نمونه، اندازه دسته‌ای، درختان رگرسیون ساخته‌شده، تعداد رقم‌های اعشاری و قابلیت‌های چاپ‌نشده و کنترل نشده همگی از مزایای مدل‌های M5P هستند (Khosravi et al., 2018). چهار مرحله اصلی در الگوریتم M5P وجود دارد که اولین آن‌ها شامل ساخت درخت با تقسیم فضای ورودی به چندین زیر فضا است. یک معیار در طول تقسیم برای رسیدگی به تغییرات درون زیر فضایی از ریشه درخت تا گره استفاده می‌شود. تغییرپذیری فضای فرعی با محاسبه انحراف استاندارد مقادیری که به گره موردنظر رسیده‌اند تعیین می‌شود. درخت با استفاده از ضریب کاهش انحراف استاندارد ساخته‌شده است. این رویکرد به‌طور قابل‌توجهی خطاهای مورد انتظار در گره را به حداقل می‌رساند (Quinlan, 1992; Khosravi et al., 2018).

جنگل تصادفی (RF)

جنگل تصادفی، اولین بار توسط Breiman و همکاران تعیین شد (Breiman, 2001). یک روش مجموعه‌ای برای ایجاد مدل‌های پیش‌بینی‌کننده برای کارهای طبقه‌بندی و رگرسیون است. این راهی برای ترکیب مدل‌های پایه کم‌تر پیش‌بینی‌کننده برای تولید مدل‌های پیش‌بینی بهتر است. مدل‌های RF به دلیل ماهیت ساده، مفروضات پایین و عملکرد بالا، به‌طور گسترده‌ای در یادگیری ماشین (ML)[1] استفاده می‌شوند. اصطلاح «جنگل» به مجموعه‌ای از

[1] Machine learning

درختان تصمیم‌گیری اشاره می‌کند که به‌تنهایی طبقه‌بندی کننده «ضعیف» هستند. یک جنگل رگرسیون قدرت پیش‌بینی یک درخت رگرسیون منفرد را ندارد. درجایی که یک درخت تنها به یک معیار تقسیم شود، به مجموعه داده آموزشی بسیار حساس است. حتی تغییرات کوچک در مجموعه داده و معیار تقسیم می‌تواند ساختارهای مختلف درخت را ترسیم کند و توضیحات متفاوتی ارائه دهد؛ بنابراین، مدل‌های RF متغیرها را بر اساس اهمیت آن‌ها برای دست‌یابی به بهترین مدل RF طبقه‌بندی می‌کنند (Breiman, 1996, 2001).

معیارهای ارزیابی مدل

در این پژوهش برای ارزیابی عملکرد سناریوهای مختلف تعریف‌شده از پارامترهای آماری ضریب همبستگی (R) (Asuero et al., 2006)، میانگین خطای مطلق (MAE)[1] (Coyle & Lin, 1988)، ضریب نش‌ساتکلیف (NS)[2] (Model, 1997) و شاخص توافق ویلموت (WI)[3] (Willmott et al., 1985) استفاده‌شده است.

$$R = \frac{\left(\sum_{i=1}^{N} Oi Pi - \frac{1}{N}\sum_{i=1}^{N} Oi \sum_{i=1}^{N} Pi\right)}{\left(\left(\sum_{i=1}^{N} Oi^2 - \frac{1}{N}\left(\sum_{i=1}^{N} Oi\right)^2\right)\left(\sum_{i=1}^{N} Pi^2 - \frac{1}{N}\left(\sum_{i=1}^{N} Pi\right)^2\right)\right)} \quad (۲)$$

$$MAE = \frac{\sum_{i=1}^{n} |Pi - Oi|}{n} \quad (۳)$$

$$NS = 1 - \left[\frac{\sum_{i=1}^{N} (Oi - Pi)^2}{\sum_{i=1}^{N} (Oi - \overline{O}i)^2}\right] \quad (۴)$$

$$WI = 1 - \left[\frac{\sum_{i=1}^{N} (Oi - Pi)^2}{\sum_{i=1}^{N} (|Pi - \overline{O}i| + |Oi - \overline{O}i|)^2}\right] \quad (۵)$$

نتایج و بحث

در این مطالعه، هفت ترکیب مختلف از متغیرهای پیش‌بینی کننده پارامترهای میانگین دما (T_{avg})، حداکثر دما (T_{max})، رطوبت نسبی متوسط (RH_{avg})، حداکثر رطوبت نسبی (RH_{max})، سرعت باد متوسط (U_{2ave})، حداکثر سرعت باد (U_{2max}) و ساعات آفتابی (SSH)، به‌عنوان ورودی مدل‌ها در جدول (۱) در نظر گرفته شد. در این بخش، الگوریتم‌های Kstar، LWL، M5P و RF در شهرستان سردشت بر روی هفت مجموعه داده برای پیش‌بینی مقادیر بارش روزانه استفاده‌شده است. تطابق بین مقادیر اندازه‌گیری شده و پیش‌بینی‌شده بارش در جدول (۲) ازنظر ضریب همبستگی، میانگین خطای مطلق، ضریب نش‌ساتکلیف و شاخص توافق ویلموت، طی مراحل اعتبارسنجی گزارش‌شده است.

مقایسه نتایج ارائه‌شده در جدول (۲) نشان می‌دهد که در سناریوی اول، M5P1 با داشتن کم‌ترین میانگین خطای مطلق (۲/۳۸۸ میلی‌متر بر روز)، بیش‌ترین ضریب نش‌ساتکلیف و ویلموت به ترتیب (۰/۰۶۸ و ۰/۳۸۳) بهترین مدل در این سناریو است. از میان الگوریتم‌های تنبل نیز الگوریتم Kstar1 با داشتن بیش‌ترین ضریب همبستگی (۰/۲۸۵)، عملکرد بهتری نسبت به مدل LWL1 از خود نشان داد. با اضافه شدن حداکثر دما به سناریوی اول، میزان خطای همه مدل‌ها به‌جز مدل RF کاهش‌یافته است. باوجوداین، بیش‌ترین کاهش خطا با مقدار شش درصد، مربوط به مدل M5P بوده است. لذا مدل M5P2 با داشتن کم‌ترین خطا (۲/۲۴۷ میلی‌متر بر روز)، نسبت به سایر مدل‌ها عملکرد بهتری داشت. با افزوده شدن پارامتر رطوبت نسبی متوسط به سناریوی دوم، دقت همه مدل‌ها افزایش‌یافته است. به‌طوری‌که بیش‌ترین کاهش خطا با مقدار ۳۱/۳ درصد، مربوط به مدل kstar و کم‌ترین کاهش خطا با مقدار ۲۵/۷۲ درصد، مربوط به مدل M5P بوده است. این بهبود ناشی از تأثیر مهم پارامتر رطوبت نسبی متوسط در پیش‌بینی بارش است. با افزودن پارامتر حداکثر رطوبت نسبی به سناریوی سوم، میزان خطای همه مدل‌ها کاهش‌یافته و ضریب همبستگی، ضریب نش ساتکلیف و شاخص توافق ویلموت نیز افزایش‌یافته است. به‌طوری‌که بیش‌ترین میزان کاهش خطا با مقدار چهار درصد، مربوط به مدل RF و کم‌ترین میزان کاهش با مقدار ۱/۵ درصد، مربوط به مدل M5P بوده است. با افزودن پارامتر سرعت باد متوسط به سناریوی چهارم، میزان خطای همه مدل‌ها به‌جز LWL کاهش‌یافته و ضریب همبستگی، ضریب

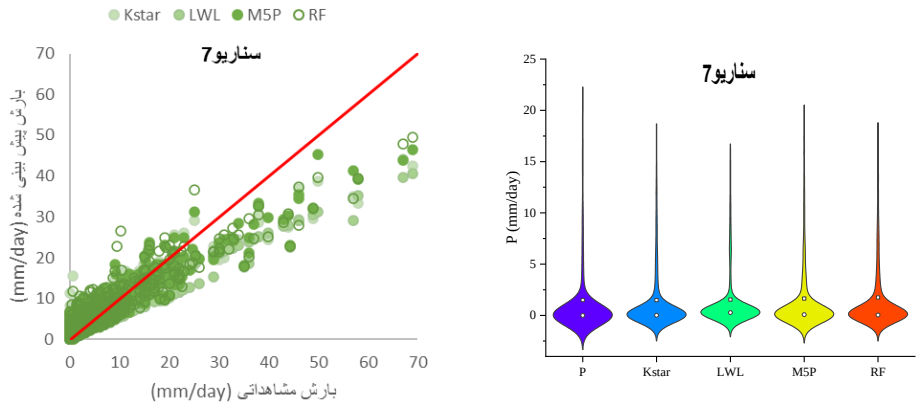
[1] Mean Absolute Error
[2] Nash-Sutcliffe coefficient
[3] Willmott's index of agreement

نش‌ساتکلیف و شاخص توافق ویلموت نیز به‌جز LWL افزایش‌یافته است. در این سناریو بیش‌ترین میزان کاهش خطا با مقدار ۲۱/۴ درصد، مربوط به مدل RF بوده است. با افزودن پارامتر حداکثر سرعت باد به سناریوی پنجم، همانند سناریوی قبلی میزان خطای همه مدل‌ها به‌جز LWL کاهش‌یافته است. در این سناریو بیش‌ترین میزان کاهش خطا با مقدار ۹/۲ درصد، مربوط به مدل M5P بوده است. با افزودن پارامتر حداکثر ساعات آفتابی به سناریوی ششم دقت تمامی مدل‌ها، با کاهش میزان خطا، افزایش‌یافته است. در این سناریو بیش‌ترین میزان کاهش خطا با مقدار سه درصد، مربوط به مدل kstar و کم‌ترین میزان کاهش با مقدار یک درصد، مربوط به مدل M5P بوده است. درنهایت مشاهده شد که با بیش‌تر شدن پارامترهای ورودی مدل‌ها، میانگین خطای مطلق کاهش می‌یابد، به‌طوری‌که درنهایت مدل درختی M5P7 با داشتن بیش‌ترین ضریب همبستگی (۰/۷۳۴)، بیش‌ترین ضریب نش‌ساتکلیف و ویلموت به‌ترتیب (۰/۵۳۶ و ۰/۸۲۷)، بهترین عملکرد را در بین تمام سناریوها دارا بوده و این مدل برآوردهای دقیق‌تری از مقدار بارش روزانه را ارائه کرد. نتایج این پژوهش با نتایج Nhu و همکاران (۲۰۲۰) که اعلام نمودند مدل درختی M5P توانایی بالایی در پیش‌بینی سطح آب روزانه دریاچه زربار دارد هم‌خوانی دارد. هم‌چنین نتایج Mirhashemi و همکاران (۲۰۲۰) که نشان دادند الگوریتم M5P توانایی بالایی در برآورد تبخیر و تعرق احتمالی، حداقل و حداکثر دما در ایستگاه هواشناسی ساری دارد، منطبق با یافته‌های پژوهش حاضر است.

جدول (۲): پارامترهای ارزیابی مدل‌های مورد مطالعه در دوره آزمون

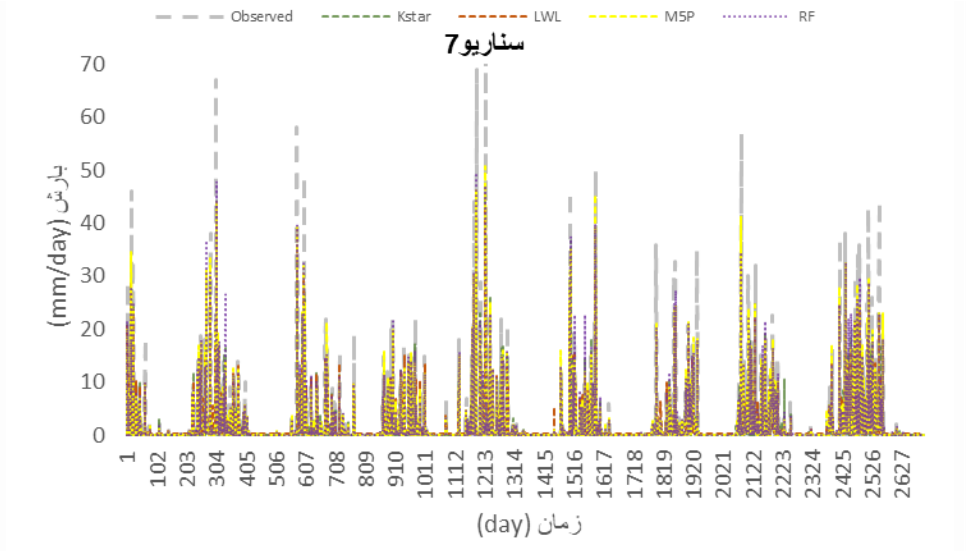
سناریو	مدل	پارامترهای آماری			
		R	MAE (mm/day)	NS	WI
سناریو ۱	Kstar1	۰/۲۸۵	۲/۳۹۲	۰/۰۷۰	۰/۳۷۸
	LWL1	۰/۲۶۸	۲/۴۲۶	۰/۰۶۰	۰/۳۶۷
	M5P1	۰/۲۸۴	۲/۳۸۸	۰/۰۶۸	۰/۳۸۳
	RF1	۰/۲۲۹	۲/۴۴۲	-۰/۰۲۲	۰/۳۸۳
سناریو ۲	Kstar2	۰/۳۰۵	۲/۳۱۳	۰/۰۸۷	۰/۳۸۹
	LWL2	۰/۲۷۵	۲/۳۹۵	۰/۰۶۶	۰/۳۶۹
	M5P2	۰/۳۵۰	۲/۲۴۷	۰/۱۱۹	۰/۴۲۹
	RF2	۰/۲۰۸	۲/۴۷۷	-۰/۳۳۱	۰/۴۰۱
سناریو ۳	Kstar3	۰/۵۸۱	۱/۶۸۷	۰/۳۳۴	۰/۶۶۶
	LWL3	۰/۵۴۳	۱/۷۹۹	۰/۲۹۰	۰/۶۵۸
	M5P3	۰/۶۴۰	۱/۷۳۵	۰/۳۸۱	۰/۷۶۸
	RF3	۰/۵۴۷	۱/۸۸۸	۰/۱۷۲	۰/۷۱۰
سناریو ۴	Kstar4	۰/۶۰۹	۱/۶۵۴	۰/۳۶۶	۰/۷۱۷
	LWL4	۰/۵۵۴	۱/۷۶۴	۰/۳۰۲	۰/۶۷۲
	M5P4	۰/۶۴۵	۱/۷۱۰	۰/۳۹۳	۰/۷۶۹
	RF4	۰/۵۷۳	۱/۸۱۳	۰/۲۲۷	۰/۷۲۹
سناریو ۵	Kstar5	۰/۶۵۰	۱/۴۳۲	۰/۴۱۰	۰/۷۱۰
	LWL5	۰/۵۴۸	۱/۷۷۷	۰/۲۹۷	۰/۶۶۰
	M5P5	۰/۶۸۶	۱/۵۴۲	۰/۴۶۵	۰/۷۸۷
	RF5	۰/۶۶۶	۱/۴۶۲	۰/۴۳۸	۰/۷۷۹
سناریو ۶	Kstar6	۰/۶۳۹	۱/۳۷۹	۰/۴۰۳	۰/۷۱۳
	LWL6	۰/۵۴۸	۱/۷۷۸	۰/۲۹۹	۰/۶۵۳
	M5P6	۰/۷۳۰	۱/۴۰۷	۰/۵۲۹	۰/۸۲۰
	RF6	۰/۶۸۸	۱/۳۹۴	۰/۴۶۹	۰/۷۹۵
سناریو ۷	Kstar7	۰/۶۵۲	۱/۳۳۸	۰/۴۲۵	۰/۷۵۱
	LWL7	۰/۵۴۳	۱/۷۴۶	۰/۲۹۴	۰/۶۴۶
	M5P7	۰/۷۳۴	۱/۳۹۴	۰/۵۳۶	۰/۸۲۷
	RF7	۰/۶۹۸	۱/۳۶۵	۰/۴۸۱	۰/۸۰۷

نمودارهای پراکنش و ویولن‌پلات مقادیر بارش روزانه (شکل ۲) نیز نشانگر عملکرد بالای مدل M5P نسبت به سایر مدل‌های KSTAR، LWL و RF در سناریوی هفتم هستند. مدل Kstar7 نیز با کم‌ترین میانگین خطای مطلق (۱/۳۳۸ میلی‌متر بر روز) در جایگاه دوم قرار می‌گیرد. در نمودار ویولن پلات دایره‌های سفیدرنگ نشان‌گر مقدار میانه و مربع‌های سفیدرنگ نشان‌گر میانگین هر مدل هستند.



شکل (۲): نمودار پراکنش و ویولن‌پلات مقادیر بارش روزانه برای بهترین سناریو

شکل (۳) تغییرات زمانی بارش روزانه را در سناریوی هفتم نشان می‌دهد. با توجه به نمودار مدل M5P دقت بالاتری نسبت به سایر مدل‌ها داشته است.



شکل (۳): نمودار تغییرات زمانی مقادیر بارش روزانه برای بهترین سناریو

نتیجه‌گیری

بارش یکی از محرک‌های اصلی در مدل‌سازی هیدرولوژیکی بوده و توزیع بارش فضایی برای درک فرآیندهای اکوهیدرولوژیکی ضروری است. لذا در این پژوهش با استفاده از روش‌های KSTAR، LWL، M5P و RF، در ایستگاه سردشت، مقادیر بارش روزانه طی دوره آماری ۲۰۲۰-۱۹۸۵ برآورد شد. داده‌های پرت از کل داده‌ها حذف‌شده و سپس نتایج به‌دست‌آمده با استفاده از پارامترهای آماری مورد مقایسه قرار گرفت و مشخص شد که روش M5P در سناریوی هفتم نتایج دقیق‌تری نسبت به سایر مدل‌ها ارائه داد. علت این امر را می‌توان به تعداد بالای پارامترهای دخیل و هم‌چنین نقش مهم پارامتر ساعات آفتابی در عملکرد مدل‌ها عنوان کرد. علاوه بر این، می‌توان نتیجه‌گیری نمود که در

حالت کلی، سناریوی هفتم مدل M5P در پیش‌بینی مقادیر بارش روزانه نتایج مناسبی ارائه داده است و برای برنامه‌ریزی‌های آبیاری و مدیریت منابع آب پیشنهاد می‌شود.

منابع

۱. پورصالحی، ف.، ع. خاشعی سیوکی و س. ر. هاشمی (۱۴۰۰) بررسی عملکرد الگوریتم جنگل تصادفی در پیش‌بینی نوسانات سطح ایستابی در مقایسه با دو مدل درخت تصمیم و شبکهٔ عصبی مصنوعی (مطالعهٔ موردی: آبخوان آزاد دشت بیرجند). اکوهیدرولوژی، (۴)۸، ۹۷۴-۹۶۱.

2. Adnan R. M., Jaafari A., Mohanavelu A., Kisi O. and Elbeltagi A. (2021) *Novel ensemble forecasting of streamflow using locally weighted learning algorithm.* Sustainability, *13(11), 5877.*
3. Asghari K. and Nasseri M. (2015) Spatial rainfall prediction using optimal features selection approaches. Hydrology Research, 46(3), 343-355.
4. Asuero A. G., Sayago A. and González A. (2006) The correlation coefficient: An overview. Critical Reviews in Analytical Chemistry, 36(1), 41-59.
5. Belachsen I., Marra F., Peleg N. and Morin E. (2017) Convective rainfall in a dry climate: relations with synoptic systems and flash-flood generation in the Dead Sea region. Hydrology and Earth System Sciences, 21(10), 5165-5180.
6. Bostan P., Heuvelink G. B. and Akyurek S. (2012) Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. International Journal of Applied Earth Observation and Geoinformation, 19, 115-126.
7. Bozorg-Haddad O., Zolghadr-Asli B., Sarzaeim P., Aboutalebi M., Chu X. and Loáiciga H. A. (2020) Evaluation of water shortage crisis in the Middle East and possible remedies. Journal of Water Supply: Research and Technology-AQUA, 69(1), 85-98.
8. Bushara NO. (2019) Weather forecasting using soft computing models: A comparative study. Journal of Applied Science, 2018 (2): 1−22.
9. Breiman L. (1996) Bagging predictors. Machine Learning, 24(2), 123-140.
10. Breiman L. (2001) Random forests. Machine Learning, 45(1), 5-32.
11. Brito S. S. B., Cunha A. P. M., Cunningham C., Alvalá R. C., Marengo J. A. and Carvalho M. A. (2018) Frequency, duration and severity of drought in the Semiarid Northeast Brazil region. International Journal of Climatology, 38(2), 517-529.
12. Cleary J. G. and Trigg L. E. (1995) K*: An instance-based learner using an entropic distance measure. In Machine Learning Proceedings, 108-114.
13. Coyle E. J. and Lin J.-H. (1988) Stack filters and the mean absolute error criterion. IEEE Transactions on Acoustics, Speech, and Signal Processing, 36(8), 1244-1254.
14. Deh-Haghi Z., Bagheri A., Fotourehchi Z. and Damalas C. A. (2020) Farmers' acceptance and willingness to pay for using treated wastewater in crop irrigation: A survey in western Iran. Agricultural Water Management, 239, 106262.
15. Dehghani M., Salehi S., Mosavi A., Nabipour N., Shamshirband S. and Ghamisi P. (2020) Spatial analysis of seasonal precipitation over Iran: Co-variation with climate indices. ISPRS International Journal of Geo-Information, 9(2), 73.
16. Di Nunno F., Granata F., Pham Q. B. and de Marinis G. (2022) *Precipitation Forecasting in Northern Bangladesh Using a Hybrid Machine Learning Model.* Sustainability, *14(5), 2663*.
17. Joshuva A., and Sugumaran V. (2020) A lazy learning approach for condition monitoring of wind turbine blade using vibration signals and histogram features. Measurement, 152, 107295.
18. He S., Wu J., Wang D. and He X. (2022) *Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest.* Chemosphere, *290, 133388.*
19. Khosravi K., Barzegar R., Miraki S., Adamowski J., Daggupati P., Alizadeh M. R., Pham B. T. and Alami M. T. (2020) Stochastic modeling of groundwater fluoride contamination: Introducing lazy learners. Groundwater, 58(5), 723-734.
20. Khosravi K., Mao L., Kisi O., Yaseen Z. M. and Shahid S. (2018) Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile. Journal of Hydrology, 567, 165-179.
21. Model U. S. (1997) MP Tripathi, RK Panda and NS Raghuwanshi. Introduction to Aquifer Analysis, 143.
22. Nhu V. H., Shahabi H., Nohani E., Shirzadi A., Al-Ansari N., Bahrami S., Miraki S., Geertsema M. and Nguyen H. (2020) Daily Water Level Prediction of Zrebar Lake (Iran): A Comparison between M5P, Random Forest, Random Tree and Reduced Error Pruning Trees Algorithms. ISPRS International Journal of Geo-Information, 9(8), 479.
23. Ostad-Ali-Askari K. (2020) The Watershed Structures in Controlling Runoff-Case Study of Sardasht Basin in IRAN. American Journal of Engineering and Applied Sciences, 13(1):72-95.

24. Pham L. T., Luo L. and Finley A. (2021) Evaluation of random forests for short-term daily streamflow forecasting in rainfall-and snowmelt-driven watersheds. Hydrology and Earth System Sciences, 25(6), 2997-3015.
25. Pour S. H. Abd Wahab A. K. and Shahid S. (2020) Spatiotemporal changes in aridity and the shift of drylands in Iran. Atmospheric Research, 233, 104704.
26. Pourghasemi H. R., Gayen A., Edalat M., Zarafshar M. and Tiefenbacher J. P. (2020) Is multi-hazard mapping effective in assessing natural hazards and integrated watershed management? Geoscience Frontiers, 11(4), 1203-1217.
27. Praveen B., Talukdar S., Mahato S., Mondal J., Sharma P., Islam A. R. M. and Rahman A. (2020) Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. Scientific reports, 10(1), 1-21.
28. Quinlan J. R. (1992) Learning with continuous classes. 5th Australian joint conference on artificial intelligence.
29. Reyes O., Cano A., Fardoun H. M. and Ventura S. (2018) A locally weighted learning method based on a data gravitation model for multi-target regression. International Journal of Computational Intelligence Systems, 11(1), 282-295.
30. Sihag P., Al-Janabi A. M. S., Alomari N. K., Ghani A. A. and Nain S. S. (2021) Evaluation of tree regression analysis for estimation of river basin discharge. Modeling Earth Systems and Environment, 7(4), 2531-2543.
31. Solomatine D. P. and Dulal K. N. (2003) Model trees as an alternative to neural networks in rainfall-runoff modelling. Hydrological Sciences Journal, 48(3), 399-411.
32. Solomatine D. P. and SIEK M. B. L. (2004) Flexible and optimal M5 model trees with applications to flow predictions. In Hydroinformatics: (In 2 Volumes, with CD-ROM) (pp. 1719-1726). World Scientific.
33. Vijayarani S. and Muthulakshmi M. (2013) Comparative analysis of bayes and lazy classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 2(8), 3118-3124.
34. Wang M. Rezaie-balf M. Naganna S. R. and Yaseen Z. M. (2021) Sourcing CHIRPS precipitation data for streamflow forecasting using Intrinsic Time-scale Decomposition based Machine Learning models. Hydrological Sciences Journal, 66(9), 1437-1456.
35. Willmott C. J. Ackleson S. G. Davis R. E. Feddema J. J. Klink, K. M. Legates D. R. O'donnell, J. and Rowe C. M. (1985). Statistics for the evaluation and comparison of models. Journal of Geophysical Research: Oceans, 90(C5), 8995-9005.
36. Wright D. B., Mantilla R. and Peters-Lidard C. D. (2017) A remote sensing-based tool for assessing rainfall-driven hazards. Environmental Modelling & Software, 90, 34-54.

[DOR: 20.1001.1.24235970.1401.10.3.1.4]

Prediction of daily precipitation of Sardasht Station using lazy algorithms and tree models

Milad Sharafi[1]*, Javad Behmanesh[2]

1. Ph.D. Student, Department of Water Engineering, Faculty of Agriculture, Urmia University, Urmia, Iran.
2. Professor, Department of Water Engineering, Faculty of Agriculture, Urmia University, Urmia, Iran.

Received: 2022/10
Accepted: 2022/12

Abstract

Due to the heterogeneous distribution of precipitation, predicting its occurrence is one of the primary and basic solutions to prevent possible disasters and damages caused by them. Considering the high amount of precipitation in Sardasht County, the people of this city turning to agriculture in recent years and not using classification models in the studied station, it is necessary to predict the daily precipitation parameter as accurately as possible. On the other hand, although the optimal performance of lazy algorithms and tree models has increased their use for predicting various hydrological phenomena, these algorithms have not been used in Sardasht County. Therefore, in this research, four models Kstar, M5P, learning algorithm with local weighting, and random forest are used to predict the daily precipitation of Sardasht Station. In this study, seven input parameters of average temperature, maximum temperature, average relative humidity, maximum relative humidity, average wind speed, maximum wind speed, and sunshine hours which were the same time as daily rainfall were used for the models. The comparison and evaluation between the input parameters showed that the sunshine hours was one of the most important input parameters, which played a significant role in the prediction accuracy of the used models. The obtained results showed that the M5P tree model had the best performance in the seventh scenario with the highest correlation coefficient (0.734 mm/day) compared to other models. In addition, the seventh scenario showed a high performance compared to the rest of the scenarios. Therefore, it can be said that increasing the input of the models has a direct relationship with their accuracy. In general, it can be said that the M5P tree model is suitable for modeling and forecasting daily rainfall in Sardasht City and it is recommended for future use.

Keywords: Modeling, Learning algorithm, Prediction, Sardasht, Tree model.

[DOR: 20.1001.1.24235970.1401.10.3.1.4]

[1] *Corresponding Author: miladsharafi1@gmail.com