Research Paper

Hybrid Random Forest-Genetic Algorithm Model with Cross Validation for Monthly Streamflow Prediction in Kashkan River Basin

Maryam Mirbeyk Sabzevari[1], **Hasan Torabi Podeh** [2*], **Mohammad Bagher Dowlatshahi** [3], **Amir Hamzeh Haghiabi** [4], **Abbas Maleki** [5]

[1] Ph.D. Candidate in Hydraulic Structures, Department of Water Engineering, Faculty of Agriculture, Lorestan University, Iran.
[2] Professor, Department of Water Engineering, Faculty of Agriculture, Lorestan University, Iran.
[3] Associate Professor, Department of Computer Engineering, Faculty of Engineering, Lorestan University, Iran.
[4] Professor, Department of Water Engineering, Faculty of Agriculture, Lorestan University, Iran.
[5] Associate Professor, Department of Water Engineering, Faculty of Agriculture, Lorestan University, Iran.

10.22125/iwe.2026.556636.1906

Received:
November 10, 2026
Accepted:
January 3, 2026
Available online:
June 18, 2026

Keywords:
Feature Selection, Genetic Algorithm, Hydrological Modeling, Random Forest, Streamflow Prediction

Abstract

This study develops and evaluates a hybrid framework for monthly streamflow prediction in the Kashkan River basin, combining a Random Forest (RF) model with a Genetic Algorithm (GA) for selecting influential features. K-fold cross validation and a structured sensitivity analysis were applied to examine the effects of removing key hydrometric stations on prediction accuracy and uncertainty. The framework aims to enhance predictions at the Kashkan-Afrine and Kashkan-Poldokhtar hydrometric stations. The dataset includes observations from three hydrometric and nine meteorological stations. The RF model, through ensemble averaging of multiple decision trees, effectively captures complex patterns and nonlinear relationships while mitigating overfitting. Results indicate that discharge from the Kakareza station exhibits a very high correlation ($R^2 = 0.97$) with downstream stations and plays a critical role in model performance. Its removal substantially reduces predictive accuracy. When all features are included, R^2 reaches 0.91 and 0.95 for Kashkan-Poldokhtar and Kashkan-Afrine, respectively, with near-zero mean prediction errors and low dispersion. Excluding Kakareza discharge lowers input correlations to approximately 0.64 and decreases R^2 to 0.72 and 0.73, while RMSE and MSE increase markedly. These findings highlight the importance of careful feature selection and the inclusion of key hydrometric stations. Overall, the proposed hybrid RF–GA framework demonstrates strong generalization capability and provides an effective and reliable approach for streamflow prediction in data-scarce river basins.

1. Introduction

Accurate streamflow prediction is vital for water management in semi-arid mountainous regions such as the Zagros, where nonlinear hydrological processes limit traditional models (Beven, 2012). Artificial intelligence and machine learning methods effectively capture these nonlinear relationships without detailed physical assumptions (Solomatine and Ostfeld, 2008). Feature selection is essential to reduce overfitting in correlated hydrological datasets (Guyon and Elisseeff, 2003), with Genetic Algorithms providing efficient global optimization.

Hybrid GA-Random Forest models enhance prediction accuracy and robustness, as RF effectively captures complex patterns and reduces variance, particularly in data-scarce basins (Cheng et al., 2006; Elshorbagy et al., 2010; Cutler et al., 2007). While most Iranian studies focus on ANNs or physical models, few integrate

* **Corresponding Author:** Hasan Torabi Podeh
Address: Department of Water Engineering, Lorestan University, Iran.
Email: torabi.ha@lu.ac.ir
Tel: 09132205169

RF with GA-based feature selection and k-fold validation. This study addresses this gap by developing a hybrid RF-GA model using multi-source data, including Kakareza station discharge, to predict monthly streamflow at Kashkan-Afrine and Kashkan-Poldokhtar stations, improving model stability, reducing overfitting, and identifying key influencing variables.

2. Materials and Methods

The study was carried out in the Kashkan River basin (7,950 km²) in southwestern Iran within the Zagros Mountains. Monthly hydro-meteorological data from three hydrometric and nine meteorological stations covering December 1969 to September 2024 were used, resulting in 633 complete records. Inputs included precipitation at time t and the previous three months, along with discharge from the Kakareza station, while monthly discharges at Kashkan–Afrine and Kashkan–Poldokhtar stations were considered as outputs, enabling multi-target downstream streamflow prediction. A Random Forest (RF) model was applied to capture nonlinear relationships and reduce overfitting through ensemble learning, using 100 trees and out-of-bag error estimation. Genetic Algorithm (GA) was employed for optimal feature selection based on a multi-objective fitness function combining prediction error and feature number. Model performance was evaluated using k-fold cross-validation and statistical criteria including R^2, RMSE, MAE, and NSE. Feature importance and model behavior were further analyzed using correlation analysis, SHAP values, and error distribution metrics.

3. Results

The results demonstrated the critical role of the upstream Kakareza station in improving downstream streamflow prediction. The hybrid RF–GA model with k-fold validation achieved high accuracy, with test-stage R^2 values of 0.91 for Kashkan–Poldokhtar and 0.95 for Kashkan–Afrine, accompanied by low errors and stable predictions. SHAP analysis identified Kakareza discharge as the dominant predictor, while precipitation and temporal variables contributed to capturing seasonal patterns. Error analyses showed limited dispersion and few outliers, though peak flows exhibited higher uncertainty.

Excluding Kakareza discharge caused a marked performance decline, with test-stage R^2 dropping to 0.72–0.73 and error dispersion increasing. SHAP results indicated greater reliance on weaker meteorological predictors, leading to higher uncertainty. Overall, the findings highlight the essential role of upstream discharge data in enhancing prediction accuracy, stability, and reliability for downstream streamflow forecasting.

4. Discussion and Conclusion

This study highlighted the critical importance of high-quality upstream hydrometric data in river flow prediction. Excluding such data markedly reduces model accuracy and stability. The hybrid RF–GA framework effectively identifies key variables and enhances generalizability when input data represent basin processes. Limitations include dependence on hydrometric data and challenges in predicting extreme events. Future work could improve performance using satellite observations, deep learning, and climate scenario simulations.

5. Six important references

1) Beven, K. (2012). Rainfall-Runoff Modelling: The Primer. Wiley-Blackwell.
2) Cheng, Q. Ma, X. and Li, Y. 2006. Optimization of hydrological model parameters using genetic algorithms: Application to Xinanjiang model. Journal of Hydrology, 316(1–4).
3) Cutler, D.R., T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson and J.J. Lawler. 2007. Random forests for classification in ecology. Ecology, 88 (11): 2783-2792.
4) Elshorbagy, A., G. Corzo, S. Srinivasulu and D. Solomatine. 2010. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology–Part 1, Concepts and methodology. Hydrology and Earth System Sciences, 14(10): 1931–1941.
5) Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.
6) Solomatine, D.P. and A. Ostfeld. 2008. Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics, 10(1): 3–22.

مدل ترکیبی جنگل تصادفی-الگوریتم ژنتیک با اعتبارسنجی متقاطع برای پیش‌بینی دبی ماهانه رودخانه در حوضه آبریز کشکان

مریم میربیک سبزواری[۱]، حسن ترابی پوده[۲*]، محمد باقر دولتشاهی[۳]، امیرحمزه حقی‌آبی[۴]، عباس ملکی[۵]

تاریخ امروز:۱۴۰۴/۰۸/۱۹
تاریخ پذیرش:۱۴۰۴/۱۰/۱۳

مقاله پژوهشی

چکیده

مطالعه حاضر به توسعه و ارزیابی یک مدل ترکیبی برای پیش‌بینی جریان رودخانه در حوضه آبریز کشکان پرداخته است. نوآوری اصلی این پژوهش، ترکیب مدل جنگل تصادفی با الگوریتم ژنتیک برای انتخاب ویژگی‌های موثر، همراه با اعتبارسنجی متقاطع kبخشی و تحلیل ساختاری اثر حذف ایستگاه‌های هیدرومتری کلیدی بر دقت و عدم قطعیت پیش‌بینی است. هدف اصلی پژوهش، بهبود دقت پیش‌بینی دبی جریان در ایستگاه‌های هیدرومتری کشکان-افرینه و کشکان-پلدختر با استفاده از این مدل ترکیبی می‌باشد. داده‌های مورد استفاده شامل مشاهدات سه ایستگاه هیدرومتری و نه ایستگاه هواشناسی است. مدل جنگل تصادفی با میانگین‌گیری از پیش‌بینی‌های مجموعه‌ای از درختان تصمیم، قادر به شناسایی الگوهای پیچیده، روابط غیرخطی و جلوگیری از بیش‌برازش است. نتایج نشان داد برخی ورودی‌ها مانند دبی ایستگاه کاکارضا همبستگی بسیار بالایی (۰/۹۷) با دبی ایستگاه‌های پایین‌دست (کشکان-افرینه و کشکان-پلدختر) دارند که نقش کلیدی در افزایش دقت مدل داشته و حذف آن‌ها منجر به کاهش قابل توجه عملکرد مدل شد. در سناریوی استفاده از تمام ویژگی‌ها، R^2 برای ایستگاه کشکان-پلدختر و کشکان-افرینه به ترتیب ۰/۹۱ و ۰/۹۵ به‌دست آمد که بیانگر قدرت تعمیم‌پذیری مناسب مدل است. خطاهای پیش‌بینی دارای میانگین نزدیک به صفر و انحراف معیار پایین بودند و پراکندگی خطاها حدود صفر بود. در مقابل حذف داده دبی کاکارضا باعث کاهش همبستگی ورودی‌ها به حدود ۰/۶۴ و افت محسوس عملکرد مدل شد به طوری‌که R^2 برای کشکان-پلدختر و کشکان-افرینه به‌ترتیب به ۰/۷۲ و ۰/۷۳ کاهش یافت در حالی‌که RMSE و MSE به‌طور قابل توجهی افزایش پیدا کردند. این نتایج اهمیت انتخاب دقیق ویژگی‌ها و بهره‌گیری از ایستگاه‌های کلیدی هیدرومتری را نشان داده و ترکیب چارچوب پیشنهادی را به‌عنوان روشی کارآمد برای پیش‌بینی دبی در حوضه‌های داده محدود تایید می‌کند.

واژه‌های کلیدی: الگوریتم ژنتیک، انتخاب ویژگی، پیش‌بینی جریان، جنگل تصادفی، مدل‌سازی هیدرولوژیکی.

---

[۱] دانشجوی دکتری سازه‌های آبی، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران، رایانامه: sabzevari.ma@fa.lu.ac.ir
[۲] نویسنده مسئول، استاد گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران، رایانامه: torabi.ha@lu.ac.ir
[۳] دانشیار گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه لرستان، ایران، رایانامه: dowlatshahi.mb@lu.ac.ir
[۴] استاد گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران، رایانامه: haghiabi.a@lu.ac.ir
[۵] دانشیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران، رایانامه: maleki.a@lu.ac.ir

مقدمه

پیش‌بینی دقیق جریان رودخانه یکی از ارکان اصلی مدیریت پایدار منابع آب، طراحی سازه‌های هیدرولیکی، بهره‌برداری از سدها و کاهش ریسک ناشی از مخاطرات هیدرولوژیکی مانند سیلاب و خشکسالی است. در مناطق کوهستانی نیمه خشک مانند زاگرس، تغییرات مکانی و زمانی شدید بارش، تبخیر و تعرق و تغییرات کاربری اراضی، روابط بین عوامل موثر بر رواناب را پیچیده و غیرخطی می‌کند. این پیچیدگی‌ها باعث می‌شود که مدل‌های مفهومی و فیزیکی کلاسیک مانند ابزار فیزیکی مبتنی بر ارزیابی آب و خاک[1] (SWAT) یا مدل مفهومی بارش-رواناب سوئدی[2] (HBV) در شرایط داده-محدود و عدم قطعیت بالا با محدودیت مواجه شوند (Beven, 2012). روش‌های هوش مصنوعی و یادگیری ماشین طی دو دهه اخیر، فرصت جدیدی برای مدل‌سازی سامانه‌های هیدرولوژیکی فراهم کرده‌اند، زیرا بدون نیاز به فرضیات پیچیده فیزیکی قادر به استخراج الگوهای غیرخطی پیچیده میان متغیرهای ورودی-خروجی از داده‌های تاریخی هستند (Solomatine and Ostfeld, 2008).

انتخاب ویژگی[3] نیز بخش مهمی از فرآیند مدل‌سازی داده‌محور است زیرا داده‌های محیطی-هیدرولوژیکی معمولا شامل متغیرهای متعدد با همبستگی بالا هستند که می‌تواند به بیش برازش و کاهش دقت مدل روی داده‌های آزمون منجر شود (Guyon and Elisseeff, 2003). برای غلبه بر این مشکل، روش‌های فراکاوشی مانند الگوریتم هارمونی[4] (Geem et al., 2001)، الگوریتم ژنتیک[5] (Holland, 1975) و بهینه‌سازی ازدحام ذرات[6] (Kennedy and Eberhart, 1995) به طور گسترده برای بهینه‌سازی انتخاب ویژگی در علوم آب به کار رفته‌اند. الگوریتم ژنتیک به‌عنوان یک روش بهینه‌سازی مبتنی بر تکامل طبیعی در مسائل پیچیده هیدرولوژی به‌ویژه کالیبراسیون مدل‌ها و انتخاب پارامترهای بهینه مدل‌های یادگیری ماشین کاربرد دارد. این الگوریتم با قابلیت جستجوی سراسری معمولا از گیر افتادن در بهینه‌های محلی جلوگیری می‌کند، اما با محدودیت‌هایی مانند نرخ همگرایی پایین و زمان محاسباتی بالا مواجه است. به همین دلیل در بسیاری از مطالعات، این الگوریتم با سایر روش‌ها ترکیب شده تا دقت و کارایی مدل بهبود یابد. با توجه به محدودیت‌های الگوریتم‌های فراابتکاری مانند مشکل بیش‌برازش، صرف زمان زیاد و بهینه‌سازی محلی، روش‌های ترکیبی برای دستیابی به پیش‌بینی‌های دقیق‌تر در زمان کوتاه‌تر مورد مطالعه قرار گرفته‌اند (Worland et al., 2018; Leonard, 2019). در میان این رویکردها، مدل‌های ترکیبی[7] مانند جنگل تصادفی[8]، ماشین تقویت گرادیان[9] (GBM) و تقویت گرادیان شدید[10] (XGBoost) به دلیل توانایی کاهش واریانس و بایاس، توانایی بالایی در پیش‌بینی متغیرهای هیدرولوژیکی یافته‌اند (Friedman, 2001; Natekin and Knoll, 2013). Cheng و همکاران (۲۰۰۶) با استفاده از الگوریتم ژنتیک پارامترهای مدل هیدرولوژیکی بارش رواناب را بهینه کردند. نتایج نشان داد GA توانست دقت پیش‌بینی جریان رودخانه را افزایش داده و بیش‌برازش را کاهش دهد. Elshorbagy و همکاران (۲۰۱۰) نشان دادند که مدل‌های ترکیبی اغلب دقیق‌تر و پایدارتر از مدل‌های منفرد هستند. الگوریتم جنگل تصادفی ترکیبی از درختان طبقه‌بندی و رگرسیون بوده و معمولا عملکرد پیش‌بینی بهتری ارائه می‌دهد (Cutler et al., 2007; Vincenzi et al., 2011). علاوه بر توان پیش‌بینی قابل قبول، این روش پیش پردازش داده‌ها را ساده‌تر و استفاده از آن را تسهیل می‌کند (Li et al., 2018). در تحقیقی Danandeh Mehr و همکاران (۲۰۱۳) با استفاده از برنامه‌نویسی ژنتیک خطی[11] (LGP)

[1] Soil and Water Assessment Tool (SWAT)
[2] Hydrologiska Byråns Vattenbalansavdelning (HBV, in Swedish)
[3] Feature Selection
[4] Harmony Search (HS)
[5] Genetic Algorithm (GA)
[6] Particle Swarm Optimization (PSO)
[7] Ensemble
[8] Random Forest (RF)
[9] Gradient Boosting Machine (GBM)
[10] Extreme Gradient Boosting (XGBoost)
[11] Linear Genetic Programming (LGP)

مدلی داده محور برای پیش‌بینی جریان رودخانه توسعه دادند و با مقایسه آن با روش موجک-شبکه عصبی نشان دادند که LGP توانایی بالایی در مدل‌سازی روابط غیرخطی سری‌های زمانی جریان و بهبود دقت پیش‌بینی دارد. Ali و همکاران (۲۰۲۰) با استفاده از مدل ترکیبی -CEEMD RF-KRR که شامل تجزیه کامل توابع ذاتی[1] (CEEMD)، جنگل تصادفی (RF) و رگرسیون کرنل ریج[2] (KRR) بود در سه منطقه پاکستان بارش ماهانه را پیش‌بینی کردند. نتایج نشان داد این مدل با تجزیه سری‌های زمانی بارش و پیش‌بینی بخش‌های مختلف با الگوریتم‌های یاد شده، عملکرد بهتری نسبت به مدل‌های مقایسه‌ای دارد و دقت بالایی در پیش‌بینی بارش ارائه می‌دهد که نشان‌دهنده قابلیت بالای مدل در مدیریت منابع آب و هشدارهای زودهنگام خشکسالی و سیلاب است.

Afan و همکاران (۲۰۲۰) نشان دادند که ترکیب الگوریتم ژنتیک با شبکه عصبی تابع پایه شعاعی[3] (RBFNN) می‌تواند به‌طور موفقیت‌آمیز پارامترهای ورودی موثر در پیش‌بینی سری‌های زمانی جریان رودخانه را تعیین کند و دقت پیش‌بینی جریان رودخانه را بهبود بخشد. Vieira و همکاران (۲۰۲۱) با استفاده از الگوریتم ژنتیک برای انتخاب ویژگی‌های موثر در داده‌های ۳۵ ساله رودخانه شینگو[4] در جنگل آمازون، مدل ترکیبی GA–Linear Regression را توسعه دادند که با ضریب تعیین ۰/۹۸۸ قادر به پیش‌بینی دقیق سطح رودخانه بود و الگوریتم ژنتیک نقش موثری در شناسایی ویژگی‌های کلیدی ایفا کرد. Islam و همکاران (۲۰۲۳) از مدل رگرسیون جنگل تصادفی[5](RFR) به‌عنوان جایگزینی برای SWAT برای پیش‌بینی جریان آب در سرچشمه‌های رود ریوگرانده[6] در ایالات متحده آمریکا استفاده کردند. نتایج نشان داد مدل رگرسیون جنگل تصادفی عملکرد دقیق‌تری نسبت به SWAT ارائه کرده و توانایی خود را در پیش‌بینی جریان آب به خوبی اثبات کرد. Chaudhary و همکاران (۲۰۲۴) نشان دادند که مدل‌های یادگیری ماشین از جمله RF و XGBoost در پیش‌بینی جریان روخانه نارمادا[7] عملکرد بالایی دارند و می‌توانند ابزار کارآمدی برای مدیریت منابع آب باشند. Panda و همکاران (۲۰۲۵) یک مدل هیدرولوژیکی تعمیم یافته مبتنی بر ترکیب موجک و جنگل تصادفی بهینه شده با الگوریتم ژنتیک (WE-RF-GA) توسعه دادند که عملکرد بسیار خوبی در پیش‌بینی جریان رودخانه نشان داد. در ایران، اغلب مطالعات پیشین بر مدل‌های کلاسیک هوش مصنوعی مانند شبکه‌های عصبی مصنوعی[8] (ANN) یا مدل‌های فیزیکی تمرکز داشته‌اند. تنها تعداد کمی از پژوهش‌ها به ترکیب RF با انتخاب ویژگی فراکاوشی و ارزیابی دقیق از طریق اعتبارسنجی متقاطع k بخشی[9] پرداخته‌اند. این مدل‌ها می‌توانند پیش‌بینی جریان رودخانه را انجام دهند اما اغلب با چالش‌هایی مانند عدم پوشش کامل عدم قطعیت‌ها، انتخاب ویژگی بهینه و اعتبارسنجی محدود مواجه هستند. برای مثال، تغییرات رطوبتی و شدت بارش در برخی مدل‌های بارش-رواناب دیده نمی‌شوند و می‌توانند دقت پیش‌بینی را کاهش دهند که خود یک شکاف پژوهشی قابل توجه محسوب می‌شود. در پژوهش حاضر، با وارد کردن ایستگاه هیدرومتری کاکارضا، این عدم قطعیت‌ها پوشش داده شده و دقت پیش‌بینی بهبود یافته است. این شرایط ضرورت انجام مطالعه‌ای را آشکار می‌سازد که ضمن استفاده از داده‌های چند منبعی (بارش، دبی و متغیرهای زمانی) الگوریتم RF را با انتخاب ویژگی GA و اعتبارسنجی چندمرحله‌ای ترکیب کند تا هم پایداری و هم دقت پیش‌بینی، بهینه شود. بنابراین اهداف اصلی این مطالعه عبارتند از: ۱) توسعه یک مدل ترکیبی شامل الگوریتم جنگل تصادفی (RF) و الگوریتم ژنتیک (GA) برای پیش‌بینی دبی جریان ماهانه در ایستگاه‌های هیدرومتری کشکان-افرینه و کشکان-پلدختر ۲) ارزیابی پایداری نتایج و کاهش بیش‌برازش مدل با استفاده از اعتبارسنجی متقاطع k بخشی ۳) مقایسه عملکرد مدل در سناریوهای مختلف (با و بدون ایستگاه هیدرومتری کاکارضا).

[1] Complete Ensemble Empirical Mode Decomposition
[2] Kernel Ridge Regression
[3] Radial Basis Function Neural Network (RBFNN)
[4] Xingu
[5] Regression Random Forest (RFR)
[6] Rio Grande
[7] Narmada
[8] Artificial Neural Network (ANN)
[9] k-fold cross-validation

این رویکرد نه تنها امکان پیش‌بینی دقیق‌تر جریان رودخانه را فراهم می‌کند بلکه به تحلیل عمیق نقش متغیرهای ورودی در فرآیند مدل‌سازی نیز کمک کرده و شکاف موجود در مطالعات پیشین را پر می‌کند. این مطالعه برای اولین بار ترکیب الگوریتم جنگل تصادفی با انتخاب ویژگی مبتنی بر الگوریتم ژنتیک و اعتبارسنجی k بخشی را برای پیش‌بینی همزمان دبی در دو ایستگاه پایین‌دست به‌طور چندهدفه به کار برده است.

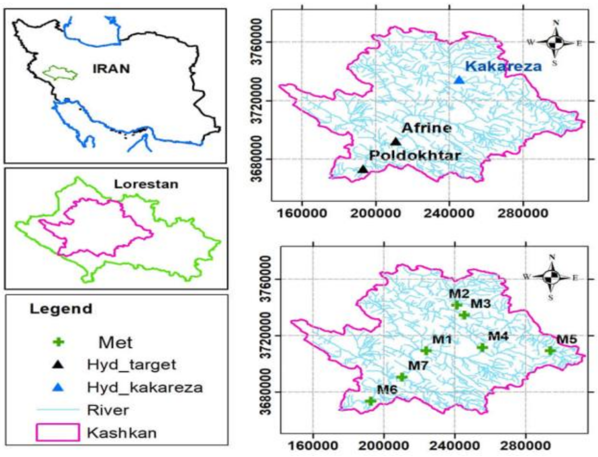
مواد و روش‌ها

منطقه مورد مطالعه

حوضه آبریز رودخانه کشکان با وسعت ۹۵۰۲/۷ کیلومترمربع در ناحیه جنوب غربی ایران واقع شده است. این حوضه از زیرحوضه‌های درجه سه حوضه آبریز کرخه بوده و ۱۲/۵ درصد از مساحت آن را شامل می‌شود. از نظر موقعیت جغرافیایی این حوضه در محدوده '۱۲ °۴۷ تا '۵۹ °۴۸ طول شرقی و '۸ °۳۳ تا '۲ °۳۴ عرض شمالی در بخش میانی کوه‌های زاگرس قرار دارد. این حوضه از لحاظ تقسیمات سیاسی در استان لرستان قرار گرفته است و ۳۳ درصد از مساحت کل استان را به خود اختصاص داده است (Lorestan Regional Water Company, 2014). در شکل ۱ موقعیت این محدوده نشان داده شده است.

داده‌ها

داده‌های مورد استفاده در این تحقیق شامل داده‌های سه ایستگاه هیدرومتری و نه ایستگاه هواشناسی می‌باشند که موقعیت این ایستگاه‌ها در شکل ۱، مشخصات ایستگاه‌های هواشناسی در جدول ۱ و مشخصات ایستگاه‌های هیدرومتری در جدول ۲ نشان داده شده است. برای مدل‌سازی از پارامترهای بارش P(t)، P(t-1)، P(t-2) و P(t-3) (به‌ترتیب میانگین بارش ماهانه در ماه جاری و سه ماه قبل) و Q(t) (میانگین دبی ماهانه در زمان t) در ایستگاه هیدرومتری کاکارضا به‌عنوان ورودی و Q(t) در دو ایستگاه هیدرومتری کشکان-افرینه و کشکان-پلدختر به‌عنوان خروجی مدل طی بازه زمانی دی ماه سال ۱۳۴۸ تا شهریور ماه ۱۴۰۳ برای حوضه آبریز کشکان استفاده گردید. اگرچه انتخاب تاخیرها بر اساس تجربه و داده‌های موجود صورت گرفت، نتایج نشان داد این مقادیر ورودی به خوبی توانستند وابستگی‌های زمانی جریان‌ها را شبیه‌سازی کنند. افزودن داده‌های ایستگاه هیدرومتری کاکارضا به‌عنوان ورودی، با هدف انتقال اطلاعات هیدرولوژیکی حوضه بالادست به مدل انجام شد.



شکل (۱): محدوده مورد مطالعه

دبی ثبت شده در این ایستگاه نمایانگر پاسخ ترکیبی فرآیندهای بارش-رواناب در بخش بالادست حوضه است و به مدل کمک می‌کند تا وابستگی زمانی و مکانی جریان‌ها را در زیرحوضه‌های مختلف بیاموزد. به‌منظور جلوگیری از

ورود عدم قطعیت اضافی و اثرگذاری بر نتایج مدل، تمام رکوردهایی که دارای داده گمشده در متغیرهای ورودی یا خروجی بودند، حذف شدند. با توجه به حذف داده‌های مفقود شده در کل داده‌های هدف و ویژگی‌ها، تعداد ۶۳۳ سری داده هم‌زمان حاصل گردید که از این داده‌ها جهت آموزش و تست مدل‌ها استفاده شد. مدل طراحی شده در این مطالعه، همزمان دبی دو ایستگاه پایین‌دست (کشکان-افرینه و کشکان-پلدختر) را پیش‌بینی می‌کند (مدل چندهدفه). این رویکرد برای کاربردهای پیش‌بینی آنلاین جریان رودخانه و امکان پیش‌بینی همزمان در چند ایستگاه مناسب است، هرچند لزوما دقت مدل را نسبت به مدل‌های تک‌هدفه افزایش نمی‌دهد.

جدول (۱): مشخصات ایستگاه‌های هواشناسی

ردیف	نام ایستگاه	ارتفاع از سطح دریا (m)	طول جغرافیایی	عرض جغرافیایی	حداکثر (mm)	میانگین (mm)	انحراف معیار	نام‌گذاری روی نقشه
۱	ویسیان	۱۰۰۸	۲۲۳۹۱۹	۳۷۰۹۳۴۶	۳۵۵	۲۳/۲	۵۳/۳۰	M1
۲	سراب صیدعلی	۱۵۷۰	۲۴۱۱۱۳	۳۷۴۱۶۵۸	۳۳۴	۲۴	۵۲/۹۴	M2
۳	کاکارضا	۱۵۳۱	۲۴۵۳۵۷	۳۷۳۴۴۱۰	۴۲۵/۴	۲۸	۶۱/۷۰	M3
۴	دره‌تنگ	۱۷۱۱	۲۴۷۸۵۸	۳۷۵۷۹۴۴	۴۵۸//۵	۲۴/۵	۵۳/۲۳	M4
۵	خرم‌آباد	۱۲۹۱	۲۵۵۵۳۲	۳۷۱۱۵۱۲	۲۸۰	۲۲	۴۷/۷۸	M5
۶	نورآباد	۱۷۸۸	۲۹۳۹۳۲	۳۷۰۹۱۴۸	۳۵۷	۲۴	۴۹/۷۱	M6
۷	کوهدشت	۱۱۹۷	۷۴۲۷۱۶	۳۷۱۳۱۳۰	۲۹۸	۱۶/۴	۴۸/۴۰	M7
۸	پلدختر	۶۶۸	۷۵۲۳۱۵	۳۶۷۱۹۵۳	۳۳۳	۱۵	۴۵/۴۶	M8
۹	افرینه	۸۲۰	۷۶۸۸۵۲	۳۶۸۹۹۱۸	۴۱۱/۵	۱۹	۵۳/۱۶	M9

جدول (۲): مشخصات ایستگاه‌های هیدرومتری

نام ایستگاه	نام رودخانه	مساحت (km^2)	ارتفاع (m)	طول جغرافیایی	عرض جغرافیایی	حداقل (m^3/s)	حداکثر (m^3/s)	میانگین (m^3/s)	انحراف معیار
کاکارضا	هررود	۱۱۵۴	۱۵۲۷	۲۴۵۳۳۳	۳۷۳۴۴۴۰	۰/۱۵۸	۲۱۳/۴۸	۴/۷۲	۱۵/۸۳
افرینه	کشکان	۶۷۰۰	۸۱۳	۷۶۹۴۳۲	۳۶۹۱۷۵۴	۰/۵۴۵	۶۵۰/۲۱	۲۳/۶۰	۴۸/۱۸
پلدختر	کشکان	۹۲۶۸	۶۷۰	۷۵۲۹۳۰	۳۶۷۲۰۱۴	۰	۸۴۹/۸۱	۲۸/۸۷	۵۹/۹۹

اجزای مدل

الگوریتم جنگل تصادفی (RF)

الگوریتم جنگل تصادفی یک الگوریتم یادگیری ماشین ترکیبی مبتنی بر درخت تصمیم است که اولین بار توسط بریمن معرفی شد (Breiman, 2001). این الگوریتم برای دو حالت دسته‌بندی و رگرسیون مورد استفاده قرار می‌گیرد و در این مطالعه برای مسئله رگرسیون و پیش‌بینی دبی جریان رودخانه به کار گرفته شده است (Cutler et al., 2012؛ Were et al., 2015). در این الگوریتم، مجموعه‌ای از درخت‌های تصمیم مستقل با استفاده از نمونه‌گیری تصادفی با جایگزینی[1] از داده‌های آموزشی ساخته می‌شوند و هر درخت با یک زیرمجموعه تصادفی از متغیرهای ورودی آموزش می‌بینند. این سازوکار موجب کاهش همزمان واریانس مدل و جلوگیری از بیش‌برازش، به‌ویژه در داده‌های هیدرولوژیکی نویزی و غیرخطی می‌شود در فرآیند مدل‌سازی حاضر، هر درخت تصمیم با استفاده از بخش تصادفی از داده‌ها آموزش داده شد و داده‌هایی که در فرآیند آموزش هر درخت استفاده نشدند، به‌عنوان نمونه‌های خارج

[1] Bootstrap sampling

از کیسه[1] (OOB) برای برآورد خطای تعمیم مدل مورد استفاده قرار گرفتند (Al-Abadi et al., 2016). این رویکرد امکان ارزیابی درونی عملکرد مدل را بدون نیاز به مجموعه آزمون مستقل فراهم می‌کند. دو پارامتر کلیدی در پیاده‌سازی مدل جنگل تصادفی شامل تعداد درخت‌ها (ntree)و تعداد متغیرهای در نظر گرفته شده در هر انشعاب درخت (mtree) است. در این مطالعه، مقدار ntree برابر با ۱۰۰ انتخاب شد تا پایداری پیش‌بینی‌ها تضمین شود، در حالی که انتخاب mtree مطابق تنظیمات استاندارد کتابخانه مورد استفاده انجام گرفت. هدف از تنظیم این پارامترها، دستیابی به تعادل مناسب بین دقت پیش‌بینی و هزینه محاسباتی و همچنین حداقل‌سازی خطای تعمیم مدل بوده است (Liaw and Wiener, 2002). در این پژوهش، عمق بیشینه درخت‌ها محدود نشده است تا مدل بتواند روابط غیرخطی و پیچیده بین متغیرهای هیدرولوژیکی ورودی و دبی خروجی را به‌صورت داده محور استخراج کند. با این حال، ماهیت تجمیعی جنگل تصادفی و میانگین‌گیری خروجی درخت‌ها مانع از بیش‌برازش می‌شود. خروجی نهایی مدل به‌صورت میانگین پیش‌بینی‌های تمامی درخت‌ها محاسبه گردید که این امر منجر به افزایش پایداری و قابلیت تعمیم نتایج شد. مقادیر پارامترهای مورد استفاده در مدل جنگل تصادفی در جدول ۳ ارائه شده است. مقادیر پیش‌فرض به تنظیماتی اطلاق می‌شود که توسط توسعه‌دهندگان کتابخانه، بر اساس آزمون‌های گسترده و کاربردهای عمومی، به‌عنوان تنظیمات استاندارد و قابل اعتماد پیشنهاد شده‌اند.

نام پارامتر	مقدار
تعداد درخت‌ها	۱۰۰
بیشینه عمق هر درخت	پیش‌فرض
حداقل تعداد نمونه لازم برای تقسیم یک گره داخلی	۲ (پیش‌فرض)
حداقل تعداد نمونه لازم در گره برگ	۱ (پیش‌فرض)
تعداد بذر تصادفی برای تضمین بازتولیدپذیری نتایج	۴۲

بهینه‌سازی انتخاب ویژگی با استفاده از الگوریتم ژنتیک

در این مطالعه به‌منظور انتخاب بهینه ویژگی‌های ورودی و کاهش پیچیدگی مدل، از الگوریتم ژنتیک به‌عنوان یک روش فراابتکاری استفاده شد. مسئله انتخاب ویژگی به‌صورت یک مسئله بهینه‌سازی تعریف گردید که در آن هدف، کمینه‌سازی هم‌زمان خطای پیش‌بینی مدل و تعداد ویژگی‌های انتخاب شده است. در این چارچوب، هر کروموزوم به‌صورت یک بردار دودویی تعریف شد که هر ژن بیانگر انتخاب (۱) یا حذف (۰) یک ویژگی ورودی می‌باشد.

فرمول کلی تابع هدف الگوریتم ژنتیک به صورت رابطه ۱ است.

$$\text{Fitness Value} = \text{MSE} * (1 + \beta * N_f) \quad (۱)$$

که MSE: میانگین مربعات خطا، N_f: تعداد ویژگی‌های انتخاب شده و β یک ضریب تنظیم‌کننده در تابع برازندگی است که به‌منظور ایجاد تعادل بین دقت پیش‌بینی و پیچیدگی مدل در فرآیند انتخاب ویژگی به کار گرفته شد. طراحی تابع برازندگی و استفاده از ضرایب کنترلی متناسب با اهداف مسئله، رویکردی رایج در الگوریتم‌های ژنتیک

[1] Out of Bag (OOB)

است (Holland, 1975). مقدار $\beta = 0.01$ بر اساس آزمون و خطا و تحلیل حساسیت انتخاب شد، به‌گونه‌ای که مقادیر بزرگ‌تر منجر به حذف بیش از حد ویژگی‌ها و افت دقت و مقادیر کوچک‌تر باعث افزایش پیچیدگی مدل بدون بهبود معنادار دقت شدند. مقدار میانگین مربعات خطا مطابق رابطه ۲ است:

$$\mathrm{MSE} = \frac{1}{n}\sum(y_i - \bar{y}_i)^2 \qquad (2)$$

که y_i مقدار مشاهده شده و $\bar{y}_i$ مقدار پیش‌بینی‌شده متغیر هدف است. در فرآیند بهینه‌سازی، برای هر کروموزوم، ابتدا زیرمجموعه ویژگی‌های انتخاب شده استخراج و سپس مدل جنگل تصادفی با همان ویژگی‌ها آموزش داده شد. مقدار MSE حاصل از پیش‌بینی مدل به‌عنوان معیار ارزیابی عملکرد استفاده و مقدار تابع هدف محاسبه گردید. الگوریتم ژنتیک با استفاده از عملگرهای انتخاب، تلاقی و جهش، جمعیت اولیه را در نسل‌های متوالی تکامل داد تا مقدار تابع هدف کمینه شود. در نهایت، کروموزومی که کمترین مقدار تابع برازش را داشت به‌عنوان مجموعه ویژگی‌های بهینه انتخاب گردید.

جدول (۴): پارامترهای الگوریتم ژنتیک

نام پارامتر	مقدار	توضیح
اندازه جمعیت	۵۰	تعداد کروموزوم‌ها در هر نسل
معیار توقف	۳۰	معیار توقف الگوریتم پس از ۳۰ نسل
نرخ جهش	۰/۰۳ (۳٪)	احتمال تغییر تصادفی ژن‌ها برای ایجاد تنوع
احتمال تقاطع	۰/۸ (۸۰٪)	احتمال انجام عمل ترکیب بین والدین

اعتبارسنجی متقاطع k بخشی

یک تکنیک ارزیابی مدل است که داده‌ها را به چند بخش تقسیم می‌کند تا از بیش برازش جلوگیری کند و تخمینی پایدار از عملکرد مدل ارائه دهد. در این روش داده‌های آموزشی به k زیرمجموعه مساوی تقسیم می‌شوند، سپس در هر تکرار، یکی از زیرمجموعه‌ها به‌عنوان داده آزمون و بقیه به‌عنوان داده آموزش استفاده می‌شوند. این فرآیند k بار تکرار می‌شود تا هر زیرمجموعه دقیقا یک بار به‌عنوان داده آزمون مورد استفاده قرار گیرد. در نهایت، میانگین معیارهای ارزیابی از تمام تکرارها به‌عنوان معیار نهایی عملکرد مدل گزارش می‌شود. این روش به ویژه زمانی مفید است که حجم داده محدود باشد زیرا از تمام داده‌ها هم برای آموزش و هم برای ارزیابی بهینه استفاده می‌کند و نتایج قابل اعتمادتری نسبت به روش‌های تقسیم ساده آموزش و تست ارائه می‌دهد. در این مطالعه، مقدار k برابر با ۵ در نظر گرفته شد. این مقدار، تعادل مناسبی بین دقت برآورد تعمیم‌پذیری مدل و زمان محاسباتی ایجاد می‌کند و برای حجم داده موجود (حدود ۶۳۳ نمونه) مناسب است. انتخاب k=۵ همچنین در مطالعات هیدرولوژی رایج بوده و عملکرد قابل اعتمادی ارائه می‌دهد.

بهینه‌سازی با الگوریتم ژنتیک

در این پژوهش، الگوریتم ژنتیک به‌منظور انتخاب بهینه ویژگی‌های ورودی مدل جنگل تصادفی به‌کار گرفته شد. هر کروموزوم به‌صورت یک بردار دودویی تعریف شد که در آن مقدار ۱ نشان‌دهنده انتخاب ویژگی و مقدار ۰ بیانگر حذف آن است. جمعیت اولیه به‌صورت تصادفی تولید و شایستگی هر کروموزوم با آموزش مدل RF و محاسبه تابع هدف ارزیابی شد. عملگر انتخاب بر اساس مقدار شایستگی انجام گرفت و سپس با اعمال تلاقی و جهش، کروموزوم‌های جدید تولید شدند تا فضای جستجو به‌طور موثر کاوش شود. این فرآیند تا رسیدن به معیار توقف ادامه یافت و در نهایت، بهترین کروموزوم به‌عنوان مجموعه ویژگی‌های بهینه برای توسعه مدل نهایی RF انتخاب شد. پارامترهای الگوریتم ژنتیک در جدول ۴ ارائه شده است.

شاخص‌های عملکرد مدل

برای ارزیابی دقت مدل از شاخص‌های استاندارد شامل ضریب تعیین[1] (R^2)، ریشه میانگین مربعات خطا[2](RMSE)، میانگین قدر مطلق خطا[3] (MAE) و ضریب نش ساتکلیف[4] (NSE) استفاده گردید (Ghorbani et al. 2016). بهترین مقادیر این معیارها به ترتیب یک،

[1] Coefficient of Determination (R^2)
[2] Root Mean Squared Error (RMSE)
[3] Mean Absolute Error (MAE)
[4] Nash–Sutcliffe Efficiency Coefficient (NSE)

صفر، صفر و یک هستند. به علاوه برای R^2، مقدار یک نشان‌دهنده دقت بالای پیش‌بینی و مقدار صفر نشان‌دهنده شکست مدل پیشنهادی در تعیین تغییرپذیری میانگین است (Nagelkerke 1991). ضریب NSE یک آمار نرمال شده است که به تعیین واریانس باقیمانده کمک می‌کند (Nash & Sutcliffe, 1970; Moriasi et al., 2007) که $-\infty < \mathrm{NSE} < 1$ می‌باشد. این ضریب در هیدرولوژی معیاری است برای سنجش تطابق بین جریان پیش‌بینی شده و جریان واقعی رودخانه. مقدار NSE نزدیک به ۱ نشان دهنده پاسخ بهینه، مقدار صفر بیانگر عملکرد برابر با میانگین داده‌ها و مقادیر منفی نشان‌دهنده عملکرد ضعیف‌تر از استفاده از میانگین است. فرمول‌های مورد نیاز برای محاسبه شاخص‌های مورد نظر در معادلات ۳ تا ۶ ارائه شده‌اند.

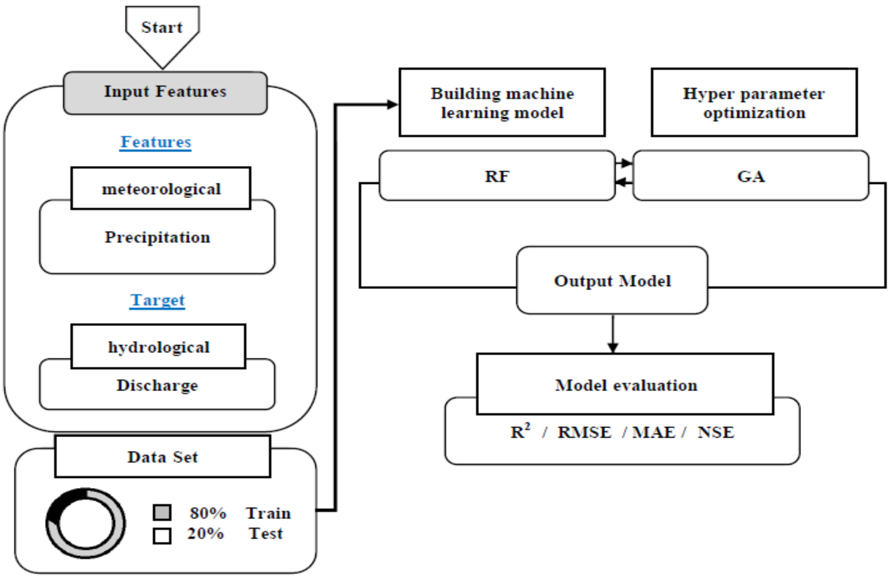
$$R^2 = \left[\frac{\sum_{i=1}^{n}(M_{oi}-\bar{M}_o)(M_{ei}-\bar{M}_e)}{\sqrt{\sum_{i=1}^{n}(M_{oi}-\bar{M}_o)^2\sum_{i=1}^{n}(M_{ei}-\bar{M}_e)^2}}\right]^2, 0 \le R^2 \le 1 \quad (۳)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_{ei}-M_{oi})^2}, 0 \le \mathrm{RMSE} \le \infty \quad (۴)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|M_{ei}-M_{oi}|, \quad 0 \le \mathrm{MAE} \le \infty \quad (۵)$$

$$\mathrm{NSE} = 1-\frac{\sum_{i=1}^{n}(M_{ei}-M_{oi})^2}{\sum_{i=1}^{n}(M_{ei}-\bar{M}_e)^2}, \quad -\infty < \mathrm{NSE} < 1 \quad (۶)$$

در روابط بالا، M_{oi} و M_{ei} به‌ترتیب مقادیر مشاهداتی و محاسباتی در گام زمانی iام، n تعداد گام‌های زمانی، $\bar{M}_o$ و $\bar{M}_e$ نیز به‌ترتیب میانگین مقادیر مشاهداتی و محاسباتی است. علاوه بر معیارهای فوق از ابزارهای تحلیل بصری مانند نقشه و ماتریس همبستگی، نمودار شاپ، هیستوگرام توزیع خطا و نمودار جعبه‌ای برای بررسی دقیق‌تر عملکرد مدل در شرایط مختلف بهره گرفته شد. در شکل ۲ نیز فلوچارت مدل ترکیبی RF-GA ارائه شده است.



شکل (۲): فلوچارت کلی مدل RF–GA

نقشه همبستگی

نقشه همبستگی[1] یکی از ابزارهای مفید در تحلیل داده‌ها است که به ما کمک می‌کند تا روابط بین ویژگی‌ها و متغیرهای هدف را به صورت بصری مشاهده کنیم. این ابزار به ویژه در یادگیری ماشین و تحلیل داده‌ها برای شناسایی ویژگی‌های مرتبط و ارزیابی کیفیت پیش‌بینی‌ها بسیار کاربردی است.

روش تفسیرپذیری ویژگی شاپ[2] (SHAP)

این روش یکی از روش‌های تفسیر مدل‌های یادگیری ماشین است که به طور خاص برای ارزیابی اهمیت ویژگی‌ها طراحی شده و بر اساس نظریه بازی‌های همکاری (نظریه شاپلی) توسعه یافته است. این روش به تحلیلگر کمک می‌کند تا بفهمد چگونه ویژگی‌های مختلف بر پیش‌بینی‌های یک مدل تاثیر می‌گذارند. شاپ اهمیت هر ویژگی را با محاسبه تاثیر آن بر پیش‌بینی مدل در مقایسه با وضعیت‌های مختلف محاسبه می‌کند. برای هر نمونه، شاپ اهمیت هر ویژگی را به صورت عددی محاسبه می‌کند که این عدد می‌تواند مثبت، منفی یا صفر باشد. فرمول شاپ برای محاسبه اهمیت ویژگی برای یک نمونه خاص به صورت فرمول ۷ بیان می‌شود.

$$\emptyset_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \left(f\left(S \cup \{i\} - f(S)\right)\right) \quad (۷)$$

$\emptyset_i$: اهمیت ویژگی i، f(S): پیش‌بینی مدل برای مجموعه ویژگی‌های S، N: مجموعه تمامی ویژگی‌ها، |S|: تعداد عناصر موجود در مجموعه S، |N|: تعداد کل ویژگی‌ها در مجموعه N، $|N| - |S| - 1$: تعداد ویژگی‌های باقی‌مانده پس از حذف ویژگی i و ویژگی‌های موجود در مجموعه S $\frac{|S|!(|N|-|S|-1)!}{|N|!}$: این عبارت نشان‌دهنده احتمال انتخاب ویژگی i در حضور ویژگی‌های موجود در مجموعه S است، عبارت $\left(f\left(S \cup \{i\} - f(S)\right)\right)$: این بخش نشان‌دهنده تفاوت در پیش‌بینی مدل f است هنگامی که ویژگی I به مجموعه S اضافه می‌شود. خواص شاپ شامل سه مورد است: جمع‌پذیری[3]، قابل توزیع بودن[4] و عدم تبعیض[5].

جمع‌پذیری: به این معنا که مجموع اهمیت ویژگی‌ها برای یک پیش‌بینی باید برابر با پیش‌بینی مدل باشد.
قابل توزیع بودن: که اگر یک ویژگی تاثیر نداشته باشد، شاپ باید اهمیت آن را صفر در نظر بگیرد.
عدم تبعیض: که اگر دو ویژگی تاثیر یکسانی بر پیش‌بینی داشته باشند باید اهمیت یکسانی نیز داشته باشند.
در تفسیر اهمیت ویژگی‌ها، عدد مثبت نشان‌دهنده این است که ویژگی به افزایش پیش‌بینی کمک می‌کند، عدد منفی یعنی ویژگی به کاهش پیش‌بینی کمک می‌کند و عدد صفر به این معنی است که ویژگی تاثیری بر پیش‌بینی نداشته است. مزایای شاپ شامل این است که یک روش تفسیر قدرتمند و علمی است که مبنای نظری قوی دارد، به خوبی با مدل‌های پیچیده مانند درخت‌های تصمیم و شبکه‌های عصبی کار می‌کند و قابلیت تفسیر جزیی و کلی دارد. از معایب شاپ می‌توان به زمان‌بر بودن محاسبه آن به ویژه برای مدل‌های بزرگ و پیچیده و نیاز به منابع محاسباتی بیشتر اشاره کرد. شاپ به تحلیلگران و دانشمندان داده کمک می‌کند تا بفهمند که چه عواملی در پیش‌بینی‌های مدل مهم هستند و چگونه می‌توانند به بهبود مدل کمک کنند. در نهایت، روش شاپ یکی از ابزارهای قوی برای تفسیر مدل‌های یادگیری ماشین است و به تحلیلگران کمک می‌کند تا اهمیت ویژگی‌ها را به طور علمی و قابل فهم ارزیابی کنند و در بهبود شفافیت و قابلیت اعتماد مدل‌ها بسیار موثر باشد.

هیستوگرام توزیع خطا

یکی از ابزارهای مهم در تحلیل عملکرد مدل‌های یادگیری ماشین است. این ابزار به ما کمک می‌کند تا درک بهتری از نحوه توزیع خطاها داشته باشیم و مشکلات احتمالی مدل را شناسایی کنیم. توزیع خطا به تفاوت بین مقادیر واقعی (مشاهده شده) و مقادیر پیش‌بینی شده توسط مدل اشاره دارد و می‌تواند اطلاعات ارزشمندی درباره عملکرد مدل ارائه دهد شامل:

[1] Correlation Heatmap
[2] SHapley Additive exPlanations (SHAP)
[3] Additivity
[4] Efficiency
[5] Symmetry

میانگین خطا[1]: نشان‌دهنده میزان پیش‌بینی درست مدل است. اگر میانگین خطا نزدیک به صفر باشد، مدل به خوبی عمل کرده است.

- انحراف معیار خطا[2]: مقدار پراکندگی خطاها را نشان می‌دهد. انحراف معیار بالاتر به معنای پیش‌بینی‌های ناپایدارتر است.
- چولگی[3]: به شکل توزیع خطا اشاره دارد و می‌تواند نشان‌دهنده این باشد که آیا مدل به طور خاص به سمت پیش‌بینی‌های بیش از حد یا کمتر از حد تمایل دارد.

بحث و نتایج

ایستگاه هیدرومتری کاکارضا به دلیل موقعیت بالادست، نقش کلیدی در کاهش عدم قطعیت و افزایش دقت پیش‌بینی دبی در ایستگاه‌های پایین‌دست کشکان-افرینه و کشکان-پلدختر دارد. برای بررسی تاثیر داده‌های کاکارضا، دو حالت الف و ب در نظر گرفته شد.

الف) عملکرد مدل با در نظر گرفتن ایستگاه هیدرومتری کاکارضا

در این مطالعه، مدل RF با استراتژی اعتبارسنجی متقاطع k بخشی و انتخاب ویژگی‌های موثر با استفاده از الگوریتم ژنتیک برای پیش‌بینی دبی جریان ماهانه در ایستگاه‌های کشکان-افرینه و کشکان-پلدختر توسعه داده شد. نتایج نشان‌دهنده دقت بالای پیش‌بینی و تعمیم‌پذیری مناسب مدل است.

عملکرد مدل در پیش‌بینی دبی ایستگاه‌های کشکان-پلدختر و کشکان-افرینه

نمودارهای شکل ۳ نشان می‌دهند که در داده‌های آموزش، مدل RF با خطای کم (MSE=106.91، RMSE=10.34، R^2=0.97) عملکردی نزدیک به ایده‌آل دارد. در داده‌های آزمون نیز با وجود افت جزیی در دقت (MSE=253.10، RMSE=15.91، R^2=0.91) نتایج حاکی از قدرت تعمیم‌پذیری مناسب مدل است. این تفاوت اندک بین آموزش و آزمون، بیانگر کنترل مناسب بیش‌برازش در مدل است.

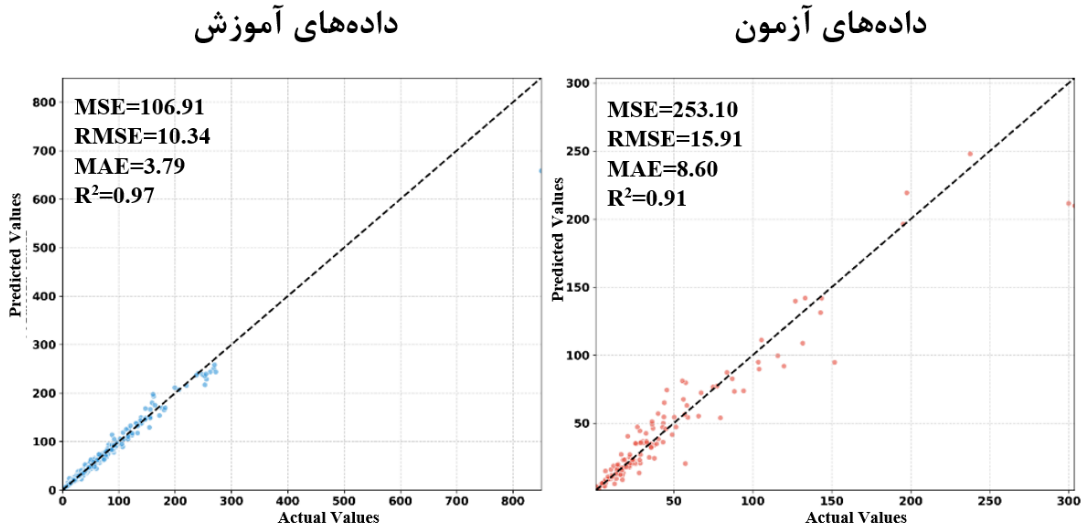
بر اساس شکل ۴، مدل در داده‌های آموزش توانسته با MSE=59.91 و R^2=0.98 دقت بالایی ارائه دهد. در داده‌های آزمون نیز با MSE=84.39، RMSE=9.19 و R^2=0.95همچنان نتایج رضایت‌بخش و قابل اعتمادی حاصل شده است. این پایداری عملکرد بین دو ایستگاه نشان‌دهنده کارایی روش انتخاب ویژگی و استفاده از اعتبارسنجی متقاطع k بخشی است. عملکرد بالای مدل ترکیبی RF-GA در پیش‌بینی دبی ماهانه حوضه کشکان، علاوه بر بهینه‌سازی عددی، دارای توجیه فیزیکی و هیدرولوژیکی نیز هست. دبی ایستگاه هیدرومتری کاکارضا، واقع در بخش بالادست حوضه، بیشترین همبستگی (۹۷٪) را با دبی ایستگاه‌های کشکان-افرینه و کشکان-پلدختر دارد و نمایانگر پاسخ ترکیبی فرآیندهای بارش و رواناب در این منطقه است. افزودن این ایستگاه به مدل، به جنگل تصادفی اجازه می‌دهد وابستگی‌های مکانی و زمانی جریان‌ها را در زیرحوضه‌های مختلف یاد بگیرد و روابط غیرخطی بین بارش، دبی بالادست و تغییرات فصلی را بهتر شناسایی کند. همچنین، استفاده از الگوریتم ژنتیک برای انتخاب ویژگی‌های موثر، باعث تمرکز مدل بر متغیرهای کلیدی شده و نقش داده‌های کم اثر را کاهش می‌دهد که این امر منجر به کاهش بیش‌برازش و افزایش پایداری و قابلیت تعمیم مدل می‌شود. بنابراین، دقت بالای پیش‌بینی و پایداری نتایج مدل، منعکس‌کننده توانایی آن در بازنمایی واقعیت‌های هیدرولوژیکی حوضه و انتقال اطلاعات حیاتی از بخش بالادست به پایین‌دست است.

ماتریس همبستگی ویژگی‌ها و متغیر هدف

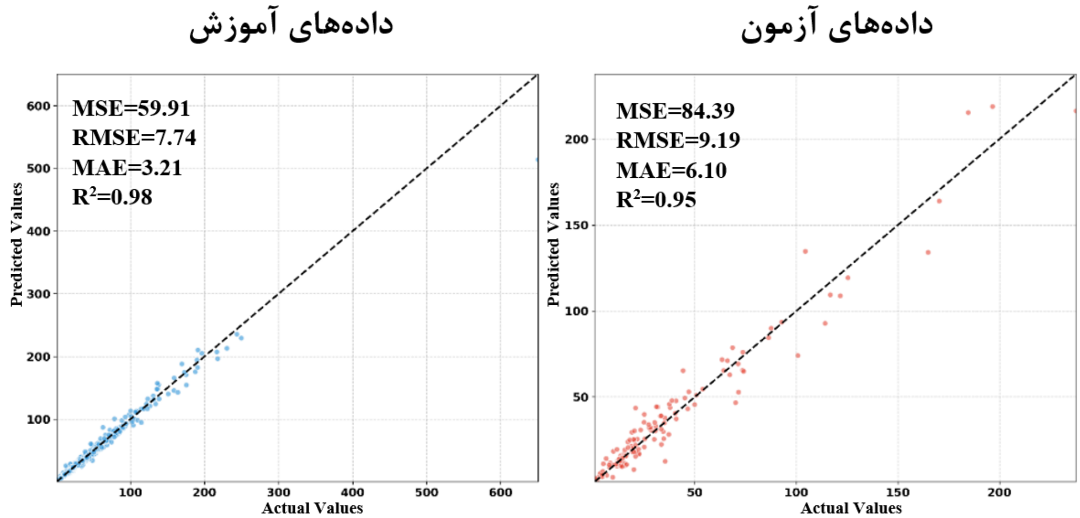
در شکل ۵، ماتریس همبستگی بین ویژگی‌های منتخب و دبی جریان در دو ایستگاه هدف نشان داده شده است. بالاترین همبستگی مثبت مربوط به دبی ایستگاه هررود کاکارضا با ضرایب ۰/۹۷ برای ایستگاه کشکان-افرینه و

[1] Mean Error
[2] Standard Deviation of Error
[3] Skewness

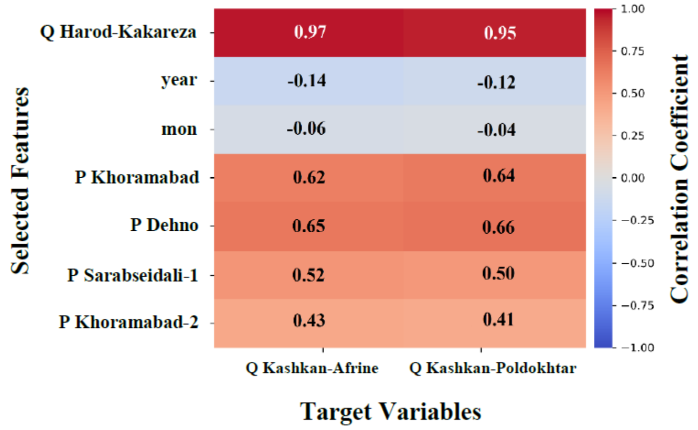
۰/۹۵ برای ایستگاه کشکان-پلدختر است که نشان می‌دهد این متغیر تاثیر قابل توجهی بر پیش‌بینی دارد. سایر پارامترها مانند بارش ایستگاه خرم‌آباد و بارش ایستگاه دهنو نیز همبستگی‌های متوسطی (بین ۰/۶۲ تا ۰/۶۶) با متغیر هدف دارند. وجود متغیرهایی با همبستگی پایین‌تر مانند سال (year) و ماه (mon) به مدل کمک می‌کنند تا الگوهای غیرخطی و فصلی را بهتر شناسایی کند.



شکل (۳): مقادیر واقعی و پیش‌بینی شده ایستگاه کشکان-پلدختر



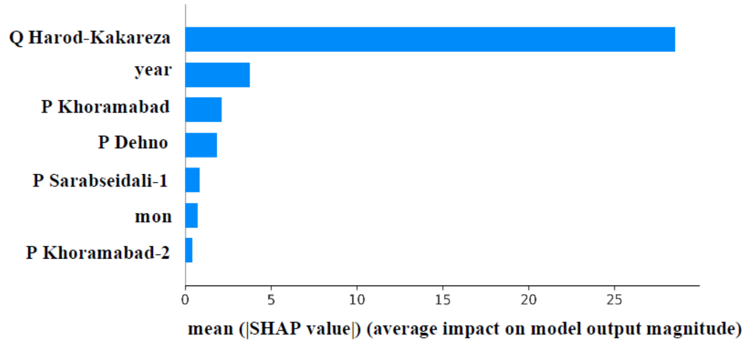
شکل (۴): مقادیر واقعی و پیش‌بینی شده ایستگاه کشکان-افرینه



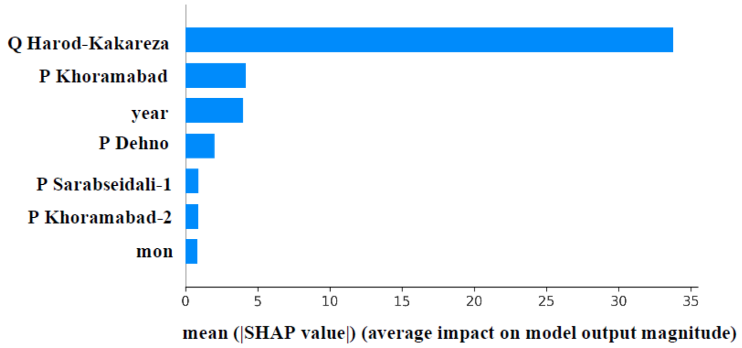
شکل (۵): ماتریس همبستگی ویژگی‌ها و متغیر هدف برای روش SHAP

نمودار شاپ برای دو ایستگاه کشکان-پلدختر و کشکان-افرینه

بر اساس نمودار ۶ و ۷، دبی ایستگاه هیدرومتری هررود کاکارضا بیشترین تاثیر را در پیش‌بینی دبی ایستگاه کشکان-افرینه و کشکان-پلدختر داشته است که بیانگر نقش کلیدی این ایستگاه در مدل‌سازی جریان رودخانه است. از منظر هیدرولوژیکی این ایستگاه بخش بالادست حوضه را پوشش می‌دهد و پاسخ ترکیبی رواناب به بارش‌های ماهانه و تجمع جریان‌ها در زیرحوضه بالادست را نشان می‌دهد. بنابراین افزودن این داده به مدل باعث می‌شود که الگوریتم جنگل تصادفی وابستگی‌های مکانی و زمانی جریان‌ها را یاد بگیرد و پایداری پیش‌بینی‌ها افزایش یابد. در مقابل، متغیرهای زمانی مانند year و mon و نیز داده‌های هواشناسی سایر ایستگاه‌ها اهمیت کمتری داشته‌اند و بیشتر نقش تکمیلی در بهبود عملکرد مدل ایفا کرده‌اند. متغیرهای زمانی به مدل کمک می‌کنند تا تغییرات فصلی و پاسخ غیرخطی رواناب به بارش‌ها را شناسایی کند. در مجموع نتایج نشان می‌دهد داده‌های هیدرومتری با همبستگی بالا، به ویژه ایستگاه بالادست، اصلی‌ترین عامل در افزایش دقت و تعمیم‌پذیری مدل هستند و تحلیل SHAPبا فرآیندهای فیزیکی حوضه همخوانی دارد.



شکل (۶): نمودار SHAP برای مدل پیش‌بینی دبی ایستگاه کشکان-افرینه (با در نظر گرفتن ایستگاه کاکارضا)



شکل (۷): نمودار SHAP برای مدل پیش‌بینی دبی ایستگاه کشکان-پلدختر (با در نظر گرفتن ایستگاه کاکارضا)

تحلیل توزیع خطاها

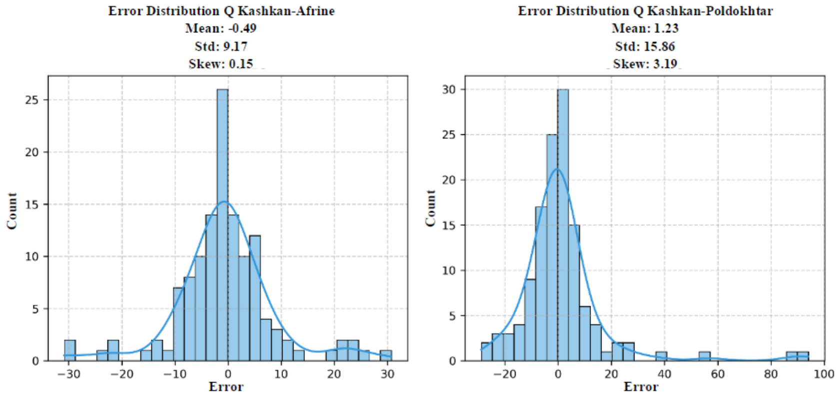
در شکل ۸ هیستوگرام توزیع خطا برای ایستگاه کشکان-افرینه، میانگین خطا برابر با ۰/۴۹- و چولگی مثبت برابر ۰/۱۵ است که نشان‌دهنده گرایش جزیی مدل به پیش‌بینی کمتر از مقادیر واقعی در برخی نقاط است. برای ایستگاه کشکان-پلدختر، میانگین خطا ۱/۲۳ و چولگی مثبت ۳/۱۹ است که بیانگر گرایش اندک به پیش‌بینی کمتر از مقدار واقعی است. در مجموع، دامنه خطاها محدود است و انحراف معیارها برای ایستگاه کشکان-افرینه و ایستگاه کشکان-پلدختر به ترتیب برابر ۹/۱۷ و ۱۵/۸۶ است که بیانگر پیش‌بینی‌های نسبتا پایدار هستند.

توزیع جعبه‌ای خطاها و نقاط پرت

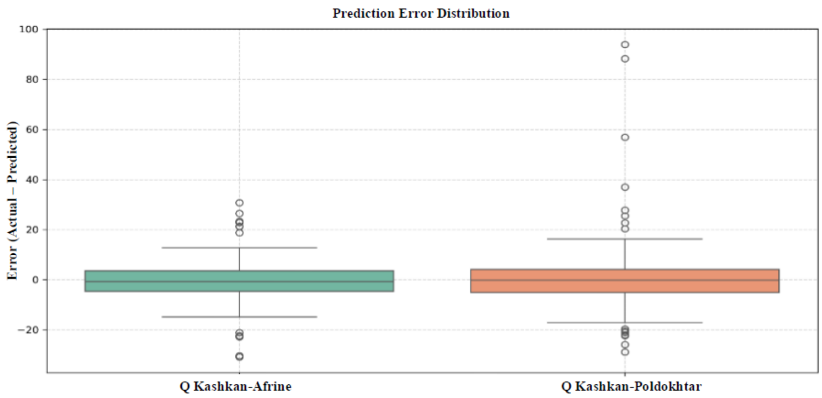
طبق شکل ۹ توزیع خطا در هر دو ایستگاه متمرکز و نزدیک به صفر است. نقاط پرت مشاهده شده نشان دهنده موارد خاص یا شرایط هیدرولوژیکی غیرمعمول هستند که مدل در پیش‌بینی آن‌ها با خطای بیشتری مواجه شده است. با این حال، پراکندگی کلی پایین و تراکم مقادیر در حوالی صفر، عملکرد مدل را تایید می‌کند. مدل RF با استفاده از انتخاب ویژگی‌های موثر و اعتبارسنجی متقاطع توانسته است روابط پیچیده و غیرخطی بین داده‌های هواشناسی-هیدرومتری و دبی جریان را به‌خوبی مدل‌سازی کند. پایداری عملکرد بین داده‌های آموزش و آزمون، مقادیر بالای ضریب تعیین و توزیع مناسب خطاها، همگی بیانگر قابلیت تعمیم‌پذیری بالا و دقت مناسب این مدل هستند. استفاده از اعتبارسنجی متقاطع k بخشی به‌طور موثر از بیش برازش جلوگیری کرده و روش انتخاب ویژگی نیز منجر به کاهش پیچیدگی مدل بدون افت محسوس در دقت شده است. بررسی هیستوگرام‌ها و نمودارهای جعبه‌ای خطا نشان می‌دهد که بیشترین مقادیر خطا عمدتا مربوط به بازه‌هایی با دبی‌های بسیار بالا است که معمولا با رخدادهای حدی مانند سیلاب‌ها همزمان هستند. از منظر هیدرولوژیکی، این موضوع قابل انتظار است زیرا در شرایط سیلابی، فرآیندهای غیرخطی مانند اشباع خاک، رواناب سریع سطحی و تغییر ناگهانی مسیر جریان نقش پررنگ‌تری پیدا می‌کنند که پیش‌بینی آن‌ها با استفاده از مدل‌های داده‌محور چالش‌برانگیزتر است. اگرچه مدل RF-GA توانسته است رفتار کلی جریان رودخانه را با دقت بالا بازنمایی کند، افزایش پراکندگی خطا در دبی‌های اوج نشان‌دهنده محدودیت مدل در شبیه‌سازی دقیق رخدادهای

حدی است. این نتایج بیانگر آن است که مدل برای کاربردهای

مدیریتی و پیش‌بینی جریان‌های نرمال و متوسط عملکرد مناسبی دارد اما در تحلیل دقیق سیلاب‌های شدید، استفاده همزمان از داده‌های هیدرولوژیکی تکمیلی یا مدل‌های فرآیندمحور می‌تواند مفید باشد.



شکل (۸): هیستوگرام توزیع خطا مدل RF-GA ایستگاه‌های کشکان-افرینه و کشکان-پلدختر



شکل (۹): نمودار جعبه‌ای توزیع خطا مدل RF-GA ایستگاه‌های کشکان-افرینه و کشکان-پلدختر

ب) عملکرد مدل بدون در نظر گرفتن ایستگاه هیدرومتری کاکارضا

عملکرد مدل در پیش‌بینی دبی ایستگاه‌های کشکان-پلدختر و کشکان-افرینه

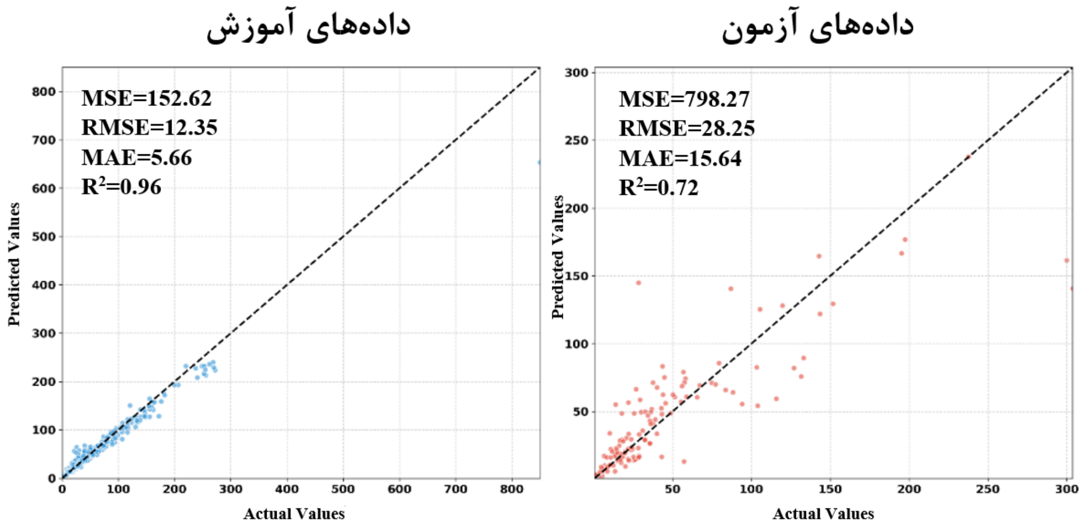
شکل ۱۰ بیانگر افت قابل توجه عملکرد نسبت به نسخه قبلی است. در داده آموزش ایستگاه کشکان-پلدختر، هنوز دقت بالا (R^2=0.96، MSE=152.62) حفظ شده است، اما در داده آزمون، ضریب تعیین به ۰/۷۲ کاهش یافته و MSE به ۷۹۸/۲۷ افزایش یافته است. RMSE و MAE نیز به ترتیب به ۲۸/۲۵ و ۱۵/۶۴ رسیده‌اند که نسبت به حالت قبل بیانگر خطای بزرگ‌تر و توان تعمیم‌پذیری ضعیف‌تر است. کاهش دقت در داده آزمون نسبت به آموزش نشان‌دهنده افزایش شکاف تعمیم مدل است که می‌تواند به دلیل از دست رفتن ورودی با قدرت پیش‌بینی بالا (ایستگاه کاکارضا) باشد.

طبق شکل ۱۱، داده‌های آموزش ایستگاه کشکان-افرینه با R^2=0.96 و MSE=102.65 عملکرد خوبی دارند، اما در آزمون، R^2 تا ۰/۷۳ افت کرده و MSE به ۴۵۵/۳۷ افزایش یافته است. این تغییرات مشابه الگوی مشاهده شده در ایستگاه کشکان-پلدختر بوده و نشان‌دهنده تاثیر مشابه حذف ایستگاه کاکارضا در هر دو ایستگاه است. افزایش RMSE به ۲۱/۳۴ و MAE به ۱۳/۱۱ حاکی از افزایش محسوس خطا در پیش‌بینی مقادیر واقعی است.

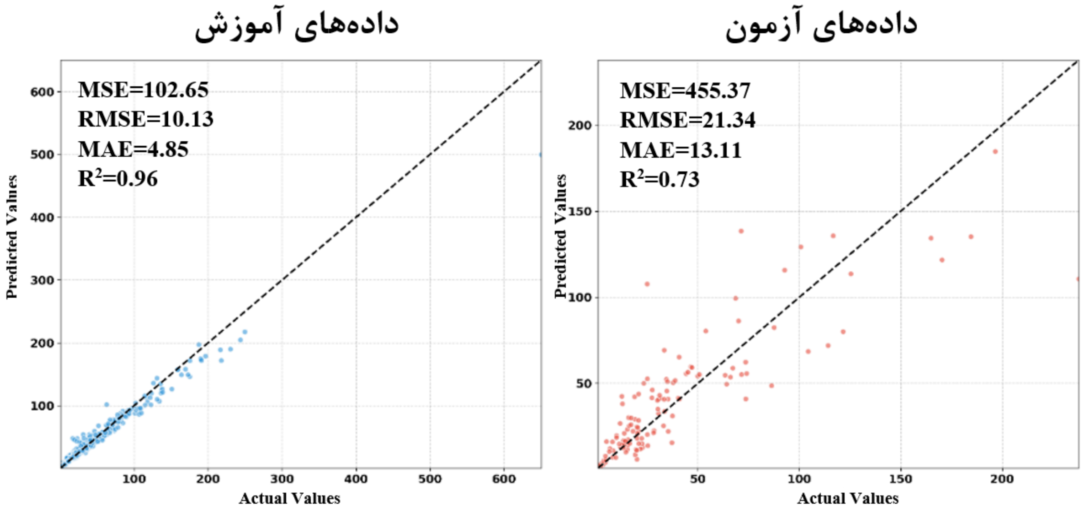
ماتریس همبستگی ویژگی‌ها و متغیر هدف

در شکل ۱۲، ضرایب همبستگی بین ویژگی‌های منتخب و دبی جریان در دو ایستگاه هدف (کشکان-افرینه و کشکان-پلدختر) نشان داده شده است. بالاترین مقدار همبستگی مثبت حدود ۰/۶۴-۰/۵۰ برای بارش ایستگاه خرم‌آباد و ایستگاه صیدعلی ثبت شده است. این در حالی است که در مجموعه داده قبلی، حضور دبی ایستگاه کاکارضا با ضریب همبستگی بسیار بالا (۰/۹۷) نقش تعیین‌کننده‌ای در بهبود مدل داشت. حذف این ویژگی باعث شده سقف مقادیر همبستگی به طور محسوس کاهش یابد و هیچ ورودی به‌تنهایی نقش غالب قبل را ایفا نکند. متغیرهای زمانی year و mon همبستگی منفی یا نزدیک به صفر دارند که نشان از تاثیر محدودشان در پیش‌بینی مستقیم دارد. لازم به ذکر است که تحلیل همبستگی مورد استفاده در این پژوهش بر اساس همبستگی خطی انجام شده است و این شاخص تنها قادر به شناسایی روابط خطی بین متغیرها می‌باشد. در حالی که مدل

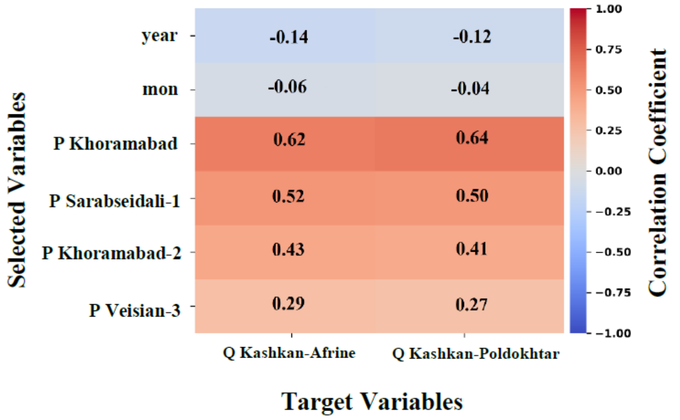
جنگل تصادفی یک مدل ذاتا غیرخطی است و می‌تواند وابستگی‌های غیرخطی و برهم‌کنش‌های پیچیده بین متغیرهای ورودی و خروجی را حتی در شرایطی با ضرایب همبستگی خطی پایین به‌خوبی استخراج کند. بنابراین، کاهش سقف همبستگی مشاهده شده در شکل ۱۲ لزوما به معنای کاهش توان پیش‌بینی مدل نیست و تحلیل همبستگی در این مطالعه صرفا به‌عنوان ابزاری توصیفی برای شناخت اولیه روابط بین متغیرها به‌کار رفته است، در حالی که تفسیر نهایی اهمیت ویژگی‌ها بر اساس عملکرد مدل و تحلیل SHAP انجام می‌شود.



شکل (۱۰): مقادیر واقعی و پیش‌بینی شده ایستگاه کشکان-پلدختر



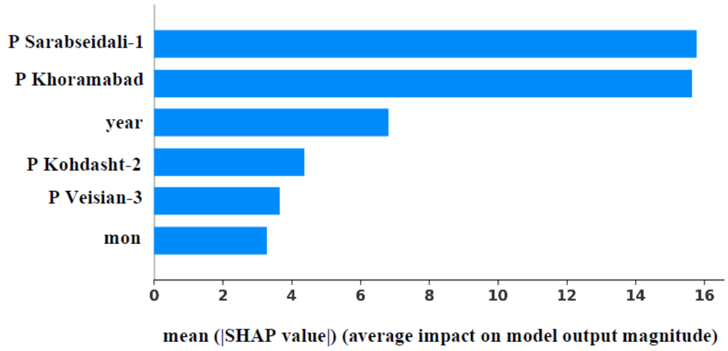
شکل (۱۱): مقادیر واقعی و پیش‌بینی شده ایستگاه کشکان-افرینه



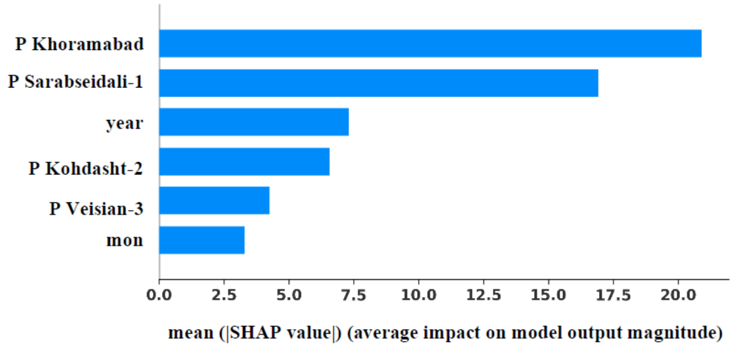
شکل (۱۲): ماتریس همبستگی ویژگی‌ها و متغیر هدف برای روش SHAP

نمودار شاپ برای دو ایستگاه کشکان-پلدختر و کشکان-افرینه

پس از حذف ایستگاه کاکارضا، نتایج نمودار SHAP نشان داد که اهمیت ویژگی‌ها به‌طور محسوسی تغییر کرده است. به‌طوری‌که ایستگاه‌های هواشناسی سراب صیدعلی و خرم‌آباد بیشترین سهم را در پیش‌بینی دبی ایستگاه کشکان-افرینه و کشکان-پلدختر داشتند (شکل ۱۳ و ۱۴). این موضوع نشان می‌دهد که با حذف متغیر کلیدی دبی کاکارضا، مدل ناچار به تکیه بیشتر بر داده‌های هواشناسی و الگوهای بارش شده است. با این حال، اتکای صرف به ایستگاه‌های هواشناسی باعث افزایش عدم قطعیت در پیش‌بینی می‌شود و افزودن حداقل یک ایستگاه هیدرومتری می‌تواند نقش موثری در بهبود پایداری و دقت مدل ایفا کند. حذف ایستگاه هیدرومتری کاکارضا موجب شد مدل بیشتر به داده‌های بارش اتکا کند که باعث افزایش عدم قطعیت پیش‌بینی‌ها و کاهش توان شبیه‌سازی پاسخ سریع حوضه به بارش‌های شدید شد. این تغییر اهمیت ویژگی‌ها پیامدهای مدیریتی دارد و نشان می‌دهد که حفظ داده‌های کلیدی بالادست برای کاهش خطا و افزایش اعتماد به پیش‌بینی‌ها ضروری است.



شکل (۱۳): نمودار SHAP برای مدل پیش‌بینی دبی ایستگاه کشکان-افرینه (بدون ایستگاه کاکارضا)



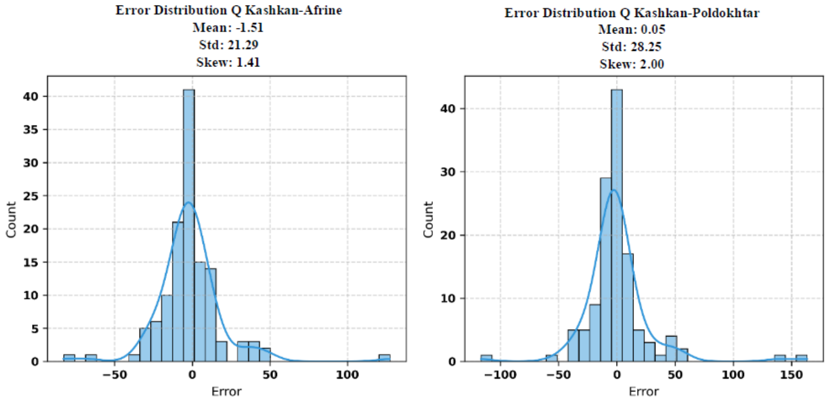
شکل (۱۴): نمودار SHAP برای مدل پیش‌بینی دبی ایستگاه کشکان-پلدختر (بدون ایستگاه کاکارضا)

هیستوگرام توزیع خطاها

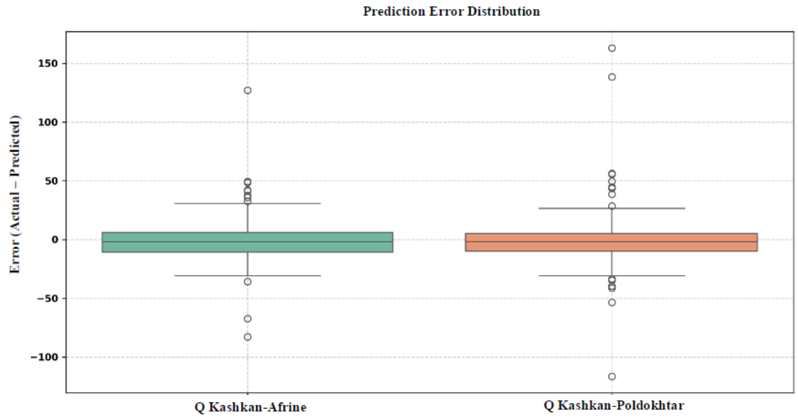
در شکل ۱۵، میانگین خطاها به مقادیر کوچک نزدیک است (۱/۵۱- برای کشکان-افرینه و ۰/۰۵ برای کشکان-پلدختر)، اما انحراف معیار آنها بالا است (۲۱/۲۹ برای کشکان-افرینه و ۲۸/۲۵ برای کشکان-پلدختر) که به معنای افزایش دامنه پیش‌بینی‌های اشتباه است. همچنین چولگی مثبت برای هر دو ایستگاه (۱/۴۱ برای کشکان-افرینه و ۲ برای کشکان-پلدختر) نشان می‌دهد که مدل بیشتر به پیش‌بینی کمتر از مقدار واقعی گرایش دارد که می‌تواند از حذف یک ورودی مهم با پیک‌های بالای همبستگی ناشی شده باشد.

نمودار جعبه‌ای توزیع خطاها

بر اساس شکل ۱۶، گستره خطاها و تعداد نقاط پرت به وضوح افزایش یافته است. مقادیر پرت در هر دو ایستگاه چه در سمت پیش‌بینی کمتر و چه بیشتر از داده واقعی، قابل توجه هستند. این الگو در مقایسه با نسخه قبلی، پراکندگی و عدم ثبات بیشتری را نشان می‌دهد.



شکل (۱۵): هیستوگرام توزیع خطا مدل RF-GA ایستگاه‌های کشکان-افرینه و کشکان-پلدختر



شکل (۱۶): نمودار جعبه‌ای توزیع خطا مدل RF-GA ایستگاه‌های کشکان-افرینه و کشکان-پلدختر

مقایسه نتایج با نسخه قبلی (با در نظر گرفتن ایستگاه کاکارضا)

قدرت همبستگی ورودی‌ها: در نسخه قبلی، وجود دبی ایستگاه کاکارضا با همبستگی حدود ۰/۹۷-۰/۹۵ با اهداف، مدل را قادر ساخت تا روابط جریان را دقیق‌تر بیاموزد. در نسخه فعلی، سقف همبستگی در حدود ۰/۶۴ است که باعث افت توان پیش‌بینی شده است.

عملکرد در داده آزمون: هر دو ایستگاه در نسخه فعلی R^2 بسیار پایین‌تری دارند (کاهش از ۰/۹۵—۰/۹۱ به ۰/۷۳—۰/۷۲) و MSE تقریبا ۳ تا ۵ برابر افزایش یافته است.

پراکندگی خطا: انحراف معیار خطا در نسخه فعلی تقریبا دو برابر شده است به‌خصوص برای پلدختر از حدود ۱۶ به حدود ۲۸، چولگی مثبت‌تر بیانگر گرایش بیشتر به کم‌برآوردی دبی است.

نقاط پرت: نمودار جعبه‌ای نسخه قبلی دامنه کمتری از خطاهای بزرگ را نشان می‌داد. نسخه فعلی نقاط پرت بیشتر و بزرگ‌تری دارد که نشان از کاهش پایداری مدل در شرایط خاص جوی-هیدرولوژیکی است.

تعمیم‌پذیری[1]: حذف کاکارضا به عنوان یک منبع داده قوی جریان، تاثیر منفی مستقیمی بر تعمیم‌پذیری مدل گذاشته به طوری‌که شکاف آموزش-آزمون در هر دو ایستگاه بیشتر شده است.

[1] Generalization

تحلیل حساسیت مدل نسبت به حذف داده‌های هیدرومتری بالادست (ایستگاه کاکارضا)

مقایسه معیارهای عملکرد نشان داد که حذف دبی ایستگاه هیدرومتری کاکارضا منجر به افت محسوس R^2 و افزایش RMSE و MAE در هر دو ایستگاه کشکان-افرینه و کشکان-پلدختر شده است. این کاهش دقت بیانگر حساسیت بالای مدل به داده‌های هیدرومتری بالادست است و از منظر هیدرولوژیکی نشان می‌دهد که جریان پایین‌دست به‌شدت تحت تاثیر پاسخ تجمعی رواناب در بخش‌های بالادست حوضه قرار دارد. در مقابل، حذف متغیرهای زمانی و داده‌های هواشناسی سایر ایستگاه‌ها افت محدودتری در عملکرد مدل ایجاد کرده که نشان‌دهنده نقش تکمیلی این متغیرها در بهبود پیش‌بینی‌ها است. این تحلیل حساسیت نشان می‌دهد که داده‌های کلیدی با همبستگی بالا، نقش غالبی در پایداری و تعمیم‌پذیری مدل ایفا می‌کنند.

نتیجه‌گیری

این مطالعه نشان داد که کیفیت و نوع داده‌های ورودی، به‌ویژه داده‌های هیدرومتری بالادست با همبستگی بالا، نقش تعیین‌کننده‌ای در موفقیت مدل‌های پیش‌بینی جریان رودخانه دارد. حذف این داده‌ها موجب کاهش دقت، تضعیف یادگیری وابستگی‌های مکانی-زمانی و کاهش پایداری مدل می‌شود. یافته‌های این پژوهش نشان می‌دهد که ترکیب الگوریتم جنگل تصادفی با انتخاب ویژگی مبتنی بر الگوریتم ژنتیک می‌تواند چارچوبی کارآمد برای شناسایی متغیرهای موثر و بهبود قابلیت تعمیم مدل‌های هیدرولوژیکی فراهم کند، مشروط بر آنکه داده‌های ورودی نماینده مناسبی از فرآیندهای فیزیکی حوضه باشند. از این رو، اتکای صرف به داده‌های هواشناسی یا متغیرهای زمانی بدون در نظر گرفتن اطلاعات هیدرومتری کلیدی، ممکن است توان پیش‌بینی مدل را محدود کند. با وجود عملکرد مناسب مدل پیشنهادی در پیش‌بینی دبی ماهانه، وابستگی آن در دسترسی به داده‌های هیدرومتری و چالش در بازتولید رخدادهای حدی از جمله محدودیت‌های این پژوهش محسوب می‌شود. در پژوهش‌های آینده می‌توان از داده‌های ماهواره‌ای، مدل‌های یادگیری عمیق و شبیه‌سازی سناریوهای اقلیمی برای بهبود دقت و افزایش پایداری و کارایی مدل در شرایط داده محدود استفاده کرد.

منابع

Afan, H.A., M.F. Allawi, A. El-Shafie, Z.M. Yaseen, A. Najah Ahmed, M. Abdul Malek, S.B. Koting, S.Q. Salih, W.H.M. Wan Mohtar, A. Sai Hin Lai, A. Sefelnasr and M. Sherif. 2020. Input attributes optimization using the feasibility of genetic nature inspired algorithm: Application of river flow forecasting. Scientific Reports, 10, Article 4684.

Ali, M., R. Prasad, Y. Xiang and Z.M. Yaseen. 2020. Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. Journal of Hydrology, 584, 124647.

Al-Abadi, A.M., and S. Shahid. 2016. Spatial mapping of artesian zone at Iraqi southern desert using a GIS-based random forest machine learning model. Modeling Earth Systems and Environment, 2(2):1-17.

Beven, K. 2012. Rainfall-Runoff Modelling: The Primer. Wiley-Blackwell.

Breiman, L. 2001. Random Forests. Machine learning, 45(1): 5-32.

Chaudhary, K. Singh, P. Kumar, R. Singh, A. and Verma, S. 2024. River stream flow prediction through advanced machine learning models for enhanced accuracy. Results in Engineering, 22, 102215.

Cheng, Q. Ma, X. and Li, Y. 2006. Optimization of hydrological model parameters using genetic algorithms: Application to Xinanjiang model. Journal of Hydrology, 316(1–4).

Cutler, D.R., T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson and J.J. Lawler. 2007. Random forests for classification in ecology. Ecology, 88 (11): 2783-2792.

Cutler A., D.R. Cutler and J.R. Stevens. 2012. Random forests In Ensemble machine learning. Springer, Boston, MA: 157-175.

Danandeh Mehr, A. Kahya, E. and Olyaie, E. 2013. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. Journal of Hydrology, 505, 240–249.

Elshorbagy, A., G. Corzo, S. Srinivasulu and D. Solomatine. 2010. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology–Part 1, Concepts and methodology. Hydrology and Earth System Sciences, 14(10): 1931–1941.

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5): 1189–1232.

Geem. Z.W., J.H. Kim and G.V. Loganathan. 2001. A new heuristic optimization algorithm: Harmony search. Simulation, 76(2): 60–68.

Ghorbani, MA., R. Khatibi, A. Geol, M.H. Fazelifard and A. Azani.2016. Modeling river discharge time series using support vector machine and artificial neural networks. Environmental Earth Sciences, 75(4): 675-685.

Guyon, I. and A. Elisseeff A. 2003. An introduction to variable and feature selection. Journal of Machine Learning Research, 3: 1157–1182.

Holland, J.H. 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press.

Islam, K.I., E. Elias, K.C. Carroll and C. Brown. 2023. Exploring Random Forest Machine Learning and Remote Sensing Data for Streamflow Prediction: An Alternative Approach to a Process-Based Hydrologic Modeling in a Snowmelt-Driven Watershed. Remote Sens, 15, 3999.

Liaw, A. and M. Wiener. 2002. Classification and Regression by Random Forest. R News 2002, 2: 18–22.

Kennedy, J. and R. Eberhart. 1995. Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, 4: 1942–1948.

Leonard, L. 2019. Using machine learning models to predict and choose meshes reordered by graph algorithms to improve execution times for hydrological modeling. Environmental Modelling & Software, 119: 84–98.

Li, X., J. Sha and Z.L Wang. 2018. Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. Environmental Science and Pollution Research, 25 (20): 19488–19498.

Lorestan Regional Water Company. 2014. Hydrological and meteorological data of Lorestan Province. Lorestan, Iran.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel and T.L. Veith .2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3): 885-900.

Nagelkerke, N.J.D. 1991. A note on a general definition of the coefficient of determination. Biometrika, 78(3) :691-692.

Nash, J.E. and J.V. Sutcliffe. 1970. River flow forecasting through conceptual models. Part I-A discussion of principles, Journal of Hydrology, 10(3): 282–290.

Natekin, A. and A. Knoll. 2013. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7, 21.

Panda, C. Panda, K. C. Singh, R. M. Singh, R.and Singh, V. P. 2025. A generalised hydrological model for streamflow prediction using wavelet ensembling. Journal of Hydrology, 655, 132883.

Solomatine, D.P. and A. Ostfeld. 2008. Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics, 10(1): 3–22.

Vieira, A.C., G. Garcia, R.E. Pabón, L.P. Cota, P. de Souza and J. Ueyama. 2021. Improving flood forecasting through feature selection by a genetic algorithm – experiments based on real data from an Amazon rainforest river. Earth Science Informatics, 14(1): 37–50.

Vincenzi, S., M. Zucchetta, P. Franzoi and M. Pellizzato. 2011. Application of a random forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecological Modelling, 222 (80): 1471–1478.

Were, K., D.T. Bui, B. Dick and B.R. Singh. 2015. A comparative assessment of support vector regression, artificial neural networks and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators, 52: 394-403.

Worland, S.C., W.H. Farmer and J.E. Kiang. 2018. Improving predictions of hydrological low flow indices in ungaged basins using machine learning. Environmental Modelling & Software, 101: 169–182.