Estimation of Missing Streamflow Data in River Using Ensemble and Machine Learning Algorithms (Case Study: Karkheh River)

Mahsa Boustani[1], Saeed Farzin[2*], Sayed-Farhad Mousavi[3]

[1] Ph.D. Candidate, Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan 35131-19111, Iran

[2*] Associate Professor, Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan 35131-19111, Iran (Corresponding author, Email: saeed.farzin@semnan.ac.ir)

[3] Professor, Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan 35131-19111, Iran

10.22125/iwe.2025.491838.1842

Received:
December 2, 2024
Accepted:
April 20, 2025
Available online:
April 25, 2025

Keywords:
Estimation of missing data, Extreme Gradient Boosting, Ensemble learning, Machine learning, Optuna

Abstract

In this study, nine ensemble and machine learning algorithms, including Xgboost, Catboost, Extra Trees, Random Forest, M5, MLP, K-NN, Decision Tree, and SVR, were employed to estimate missing daily streamflow data for the Karkheh River in southwestern Iran. To estimate the missing data at Abdolkhan and Paye-Pol stations, the daily flow data from the Hamidiyeh hydrometric station, as a neighboring station, was analyzed over a 40-year period. Hyperparameter optimization for these algorithms was carried out using the Optuna method. A thorough comparison of model performance showed that the Xgboost algorithm, by learning complex nonlinear relationships, provided the highest estimation accuracy. The results revealed that at Abdolkhan and Paye-Pol stations, Xgboost achieved the highest efficiency, with the highest coefficient of determination (R^2) values of 0.95 and 0.78, the lowest mean absolute error (MAE) values of 18.76 and 36.45, the lowest root mean square error (RMSE) values of 43.75 and 108.87, and the lowest relative root mean square error (RRMSE) values of 0.20 and 0.46, respectively. Furthermore, Taylor diagrams confirmed the superiority of the Xgboost model at both stations. These findings highlight the ability of Xgboost to overcome spatial distance challenges and handle limited data effectively. The Catboost model achieved second place among the evaluated models, with 11% and 5% lower accuracy compared to the Xgboost model at the Abdolkhan and Paye-Pol stations, respectively. The results of this study can be valuable for estimating missing flow data at other stations of this river and play a significant role in the effective management of water resources.

1. Introduction

Recent research has increasingly focused on accurately predicting river flow using various methods to optimize water resources systems and facilitate planning activities. However, a significant portion of river basins in Iran and many developing countries lack hydrometric stations. Missing data poses significant challenges to effective water resources planning and management (Sharma and Yuden., 2021).

Khampuengson and Wang (2023) have argued that missing data can lead to inaccurate analysis or even false alarms. Therefore, identifying and correcting missing values as accurately as possible is crucial.

With recent advancements, machine learning techniques and artificial intelligence-based methods have been employed in various fields to assist humans (Boustani et al., 2025; Sharififard et al., 2024). By learning from observations and input data, these models can predict new or missing data (Luna et

al., 2020). Many researchers agree that machine learning-based methods are best suited for reconstructing missing values and typically lead to significant improvements in prediction accuracy compared to traditional statistical approaches (Krysanova & White, 2015; Minns & Hall, 1996; Varga et al., 2016).

2. Materials and Methods

In this research, capabilities of nine machine learning and ensemble learning models, including Xgboost, Catboost, Extra Trees, Random Forest, ANN-MLP, K-Neighbors, CART decision tree, SVR, and M5 decision tree, are investigated for the imputation of missing data at two hydrometric stations: Abdolkhan and Paye-Pol. Additionally, the optimal hyperparameters of these models for river flow estimation are tuned using Optuna, and the best model structures are determined. Subsequently, their performance is evaluated using statistical and visual indices.

To facilitate model training and evaluation, the dataset was partitioned into training and testing subsets. Specifically, 70% of the data was allocated for model training, while the remaining 30% was reserved for testing. The selection of training and testing data was performed using Python's scikit-learn library. Python, a versatile programming language with a rich libraries and tools, was employed for all data processing and modeling tasks in this study.

3. Discussion and Conclusion

To estimate missing flow data at Abdolkhan and Paye-Pol hydrometric stations, observed flow data from the neighboring Hamidiyeh station was utilized as a feature in the training dataset. After feature selection and hyperparameter tuning, the performance of various models was compared for estimating missing river flow data at both stations.

For Abdolkhan station, during the training phase, the Mean Absolute Error (MAE) values for Catboost, Xgboost, Random Forest, Extra Trees, M5 decision tree, CART decision tree, K-NN, MLP, and SVR were 7.936, 11.09, 10.233, 12.977, 19.319, 19.265, 20.677, 25.828, and 25.380, respectively.

The Root Mean Squared Error (RMSE) for the aforementioned algorithms was calculated as 12.98, 18.21, 31.56, 31.85, 49.95, 54.73, 58.93, 65.55, and 74.77, respectively. The Relative Root Mean Squared Error (RRMSE) was correspondingly calculated as 0.06, 0.08, 0.15, 0.15, 0.24, 0.26, 0.28, 0.32, and 0.36. The R-squared values were 0.99, 0.99, 0.97, 0.97, 0.93, 0.92, 0.91, 0.89, and 0.86, respectively.

Similarly, during the training phase for Paye-Pol station, the MAE values for Xgboost, Catboost, Extra Trees, Random Forest, M5, K-NN, Decision Tree, MLP, and SVR were 17.64, 25.04, 28.63, 25.56, 38.74, 40.13, 40.39, and 39.23, respectively.

The Root Mean Squared Error (RMSE) for the aforementioned algorithms was calculated as 31.67, 61.11, 88.30, 90.86, 123.75, 125.96, 126.19, 126.27, and 130.67, respectively. The R-squared values were correspondingly calculated as 0.98, 0.92, 0.85, 0.84, 0.72, 0.71, 0.70, 0.70, and 0.68, respectively.

Based on the evaluation metrics during the training phase, accuracy of the algorithms used for both hydrometric stations deemed acceptable. To select the best algorithm, new data was used for the testing phase. In the testing phase, the results were evaluated using MAE, RMSE, RRMSE, R2, and Taylor diagrams. Results indicated that the Xgboost model, with the lowest MAE of 18.76 and the highest R2 of 0.95, performed the best among the models for estimating daily flow in the Karkheh River at the Abdolkhan hydrometric station.

At Paye- Pol station, the Catboost model exhibited a strong performance in estimating missing data, following closely behind the Xgboost model. The remaining models demonstrated moderate performance. Conversely, the SVR model consistently showed the poorest performance at both the Abdolkhan and Paye-Pol stations. At Abdolkhan, SVR yielded R2, RRMSE, RMSE, and MAE values of 0.84, 0.39, 82.93, and 27.04, respectively. At Paye-Pol, these values deteriorated to 0.70, 0.54, 127.41, and 38.52, respectively, underscoring its inferior performance compared to Catboost, Xgboost, Random Forest, Extra Trees, M5 decision tree, CART decision tree, K-NN, and MLP.

Based on the results, application of data mining techniques for estimating daily river flow data, especially in developing countries with limited access to hydrological data, can be highly beneficial for implementing effective management strategies that mitigate the negative impacts of extreme events.

4. Results

The most significant findings of this research are as follows:

1. In both Abdolkhan and Paye-Pol hydrometric stations, ensemble learning models exhibited higher accuracy compared to machine learning models.
2. Even though Paye-Pol station is located at significant distance from the neighboring Hamidiyeh station, the Xgboost model demonstrated remarkable ability to accurately estimate missing daily river flow data, outperforming models such as Catboost, Extra Trees, Random Forest, M5, MLP, K-NN, Decision Tree, CART, and SVR.
3. Although the Abdolkhan station had a larger proportion of missing data, the models performed more accurately at this station compared to Paye-Pol station. The proximity of the Abdolkhan station to the Hamidiyeh hydrometric station is believed to be a contributing factor to this result.
4. When compared to the Xgboost model, the Catboost model exhibited a 11% and 5% decrease in accuracy at the Abdolkhan and Paye-Pol stations, respectively, ranking second among the evaluated models.
5. Based on evaluation metrics such as MAE, RMSE, RRMSE, R^2, and Taylor diagrams, the SVR model exhibited the poorest performance among the nine models assessed.
6. Ensemble learning models can overcome the limitations of individual methods, such as Support Vector Regression, leading to improved accuracy in river flow estimation.
7. Models based on boosting techniques have demonstrated the ability to effectively address spatial distance challenges and manage limited data.
8. Application of the Optuna optimization method for hyperparameter tuning of machine learning models has significantly enhanced model performance.

5. Six important references

1) Khampuengson, T., & Wang, W. 2023. Novel methods for imputing missing values in water level monitoring data. Water Resources Management, 37(2), 851-878.
2) Krysanova, V., & White, M. 2015. Advances in water resources assessment with SWAT-an overview. Hydrological Sciences Journal, 60(5), 771-783.
3) Luna, A. M., Lineros, M. L., Gualda, J. E., Giráldez Cervera, J. V., & Madueño Luna, J. M. 2020. Assessing the best gap-filling technique for river stage data suitable for low-capacity processors and real-time application using IoT. Sensors, 20(21), 6354.
4) Minns, A. W., & Hall, M. J. 1996. Artificial neural networks as rainfall-runoff models. Hydrological Sciences Journal, 41(3), 399-417.
5) Sharma, V., & Yuden, K. 2021. Imputing missing data in hydrology using machine learning models. International Journal of Engineering Research and Technology, 10, 78-82.
6) Varga, M., Balogh, S., & Csukas, B. 2016. GIS based generation of dynamic hydrological and land patch simulation models for rural watershed areas. Information Processing in Agriculture, 3(1), 1-16.

Conflict of Interest

Authors declared no conflict of interest.

تخمین داده‌های مفقود جریان رودخانه با استفاده از الگوریتم‌های یادگیری جمعی و ماشینی (مطالعه موردی: رودخانه کرخه)

مهسا بوستانی[۱]، سعید فرزین[۲*]، سیدفرهاد موسوی[۳]

تاریخ ارسال:۱۴۰۳/۰۹/۱۲
تاریخ پذیرش:۱۴۰۴/۰۱/۳۱

مقاله پژوهشی برگرفته از رساله دکتری

چکیده

در این پژوهش، از ۹ الگوریتم یادگیری جمعی و ماشینی شامل الگوریتم‌های Xgboost، Catboost، Extra Trees، Random Forest، M5، MLP، K-NN، Decision Tree و SVR برای تخمین داده‌های مفقود جریان روزانه رودخانه کرخه استفاده شد. جهت برآورد داده‌های مفقود ایستگاه عبدالخان و پای پل، داده‌های جریان روزانه ایستگاه هیدرومتری حمیدیه به عنوان ایستگاه همسایه در دوره آماری ۴۰ ساله مورد بررسی قرار گرفت. بهینه‌سازی فراپارامترهای الگوریتم‌های مذکور، به روش Optuna انجام شد. مقایسه عملکرد مدل‌ها نشان داد که الگوریتم Xgboost با یادگیری روابط غیرخطی پیچیده، دقت بیشتری در تخمین داده‌های مفقود دارد. الگوریتم مذکور، در ایستگاه‌های عبدالخان و پای پل، با داشتن بیشترین مقدار ضریب تعیین (R^2) به ترتیب برابر با ۰/۹۵ و ۰/۷۸ و کمترین مقدار میانگین خطای مطلق (MAE) به ترتیب برابر با ۱۸/۷۶ و ۳۶/۴۵ بهترین عملکرد را دارد. همچنین، کمترین مقدار ریشه میانگین مربع خطاها (RMSE) برابر با ۴۳/۷۵ و ۱۰۸/۸۷ به‌دست آمد. علاوه بر این، الگوریتم Xgboost کمترین مقدار مجذور میانگین مربعات خطای نسبی (RRMSE) برابر با ۰/۲۰ و ۰/۴۶ ثبت کرد. بنابراین، الگوریتم Xgboost بیشترین کارایی را در تخمین داده‌های مفقود نسبت به بقیه مدل‌ها در هر دو ایستگاه دارد. همچنین، می‌تواند بر چالش‌های فاصله مکانی و داده‌های محدود غلبه کند. نتایج نمودار تیلور نیز حاکی از برتری مدل Xgboost در هر دو ایستگاه مذکور است. مدل Catboost نیز در ایستگاه‌های عبدالخان و پای پل به ترتیب ۱۱٪ و ۵٪ دقت کمتر از مدل Xgboostداشت و دومین جایگاه را میان مدل‌های بررسی‌شده کسب کرد. نتایج این پژوهش می‌تواند جهت تخمین جریان رودخانه در سایر ایستگاه‌های فاقد آمار مفید واقع شود.

واژه‌های کلیدی: تخمین داده‌های مفقود، تقویت گرادیان شدید، یادگیری جمعی، یادگیری ماشین، Optuna

[۱] دانشجوی دکترای تخصصی، گروه مهندسی آب و سازه‌های هیدرولیکی، دانشکده مهندسی عمران، دانشگاه سمنان، ایران mahsaboustani@semnan.ac.ir
[۲] نویسنده مسئول و دانشیار، گروه مهندسی آب و سازه‌های هیدرولیکی، دانشکده مهندسی عمران، دانشگاه سمنان، ایران saeed.farzin@semnan.ac.ir
[۳] استاد، گروه مهندسی آب و سازه‌های هیدرولیکی، دانشکده مهندسی عمران، دانشگاه سمنان، ایران fmousavi@semnan.ac.ir

مقدمه

امروزه، تحقیقات با هدف پیش‌بینی دقیق جریان رودخانه‌ها با استفاده از روش‌های مختلف به منظور بهره‌برداری بهینه از سیستم‌های منابع آب و انجام فعالیت‌های برنامه‌ریزی مجدد از اهمیت خاصی برخوردار است و همچنان در حال افزایش است (بوستانی و همکاران، ۱۳۹۸). اما متأسفانه بخش قابل توجهی از حوضه‌های آبریز ایران و بسیاری از کشورهای در حال توسعه، فاقد ایستگاه‌های هیدرومتری هستند و یا داده‌های اندازه‌گیری شده آنها بسیار محدود است، که برای اهداف برنامه‌ریزی منابع آب مناسب نیستند، زیرا پژوهشگران برای انجام تحلیل‌های دقیق به داده‌های قابل اعتماد نیاز دارند. بنابراین، مدیران منابع آب به‌منظور برنامه‌ریزی، به مدل‌های مختلف پیش‌بینی با دقت زیاد وابسته هستند (Razavi and Coulibaly, 2013).داده‌های از دست رفته به دلایل مختلف از جمله خرابی ایستگاه‌ها، حمل و نقل و مشکلات اقلیمی و خطای انسانی، مشکلات قابل توجهی را از نظر برنامه‌ریزی و مدیریت برای عملکرد بهینه منابع آب ایجاد می‌کنند. داده‌های ثبت شده از رویدادهای طبیعی، خصوصاً در کشورهای در حال توسعه، محدود است. با این مقادیر محدود، مدل‌سازی و تحلیل قابل اعتماد رفتار هیدرولوژیک بسیار دشوار است (Sharma and Yuden, 2021).

Khampuengson and Wang (2023) اظهار داشته‌اند که معمولاً داده‌های مفقود می‌تواند منجر به تجزیه و تحلیل یا حتی هشدارهای نادرست شود. بنابراین، شناسایی مقادیر از دست رفته و تصحیح آنها تا حد امکان بسیار مفید است.

امروزه با توجه به پیشرفت‌های حاصل شده، روش‌های مبتنی بر هوش مصنوعی، تحولی بزرگ در تحلیل داده‌ها، پیش‌بینی روندها و حل مسائل پیچیده ایجاد کرده است. در این راستا می‌توان به مطالعات بوستانی و همکاران (۲۰۲۵)، شریفی‌فرد و همکاران (۲۰۲۴)، احدیان و همکاران (۲۰۲۳)، باقریان و همکاران (۲۰۱۴)، احدیان (۲۰۱۶) و دریایی و همکاران (۲۰۱۰) اشاره نمود.

از مهم‌ترین ویژگی‌های الگوریتم‌های یادگیری ماشین، برخلاف روش‌های کلاسیک، عدم وابستگی به نوع داده‌های ورودی و ماهیت آنها است. از طرفی، انتخاب داده‌ها در این نوع مدل‌ها به صورت تصادفی بوده و تأثیر هرگونه عامل ناایستایی در سری داده‌ها از بین می‌رود. این امر مزیت و قابلیت مهم این مدل‌ها محسوب می‌شود (Molnar, 2020).

Venkatesan and Mahindrakar (2019) با استفاده از الگوریتم تقویت گرادیان شدید (Xgboost) سیلاب یک تا ۵ ساعته برای حوضه کولار در هند با استفاده از داده‌های بارش و رواناب ساعتی سال‌های ۱۹۸۷ تا ۱۹۸۹ را پیش‌بینی کردند. نتایج مدل با الگوریتم جنگل تصادفی (RF) و ماشین بردار پشتیبان (SVM) مقایسه گردید و بیان شد که روش Xgboost عملکرد بهتری داشته است.

Li et al. (2020) از رگرسیون شبکه الاستیک (ENR)، رگرسیون بردار پشتیبان، RF و مدل Xgboostبرای پیش‌بینی جریان ماهانه رودخانه استفاده کردند و یک مدل چندگانه اصلاح‌شده به نام استراتژی تجمیع اصلاح‌شده (MSES) را به عنوان یک روش ادغام پیشنهاد کردند. آن‌ها بیان کردند که مدل‌های RF و Xgboostپیش‌بینی عملکرد بهتری نسبت بهENR و SVRدارند.

Ni et al. (2020) یک مدل ترکیبی ایجاد کردند که از ترکیب Xgboost و مدل ترکیبی گوسی (GMM) برای تخمین جریان ماهانه استفاده می‌کرد. آن‌ها از داده‌های جریان ماهانه ایستگاه‌های کانتان و هانکوی واقع در حوضه رودخانه یانگ‌تسه برای مدل‌سازی استفاده کردند. آن‌ها این مدل را به عنوان یک گزینه برتر برای مدیریت بهینه منابع آب پیشنهاد کردند.

Latifoğlu and Canpolat (2022) با استفاده از تکنیک‌های Bagging و Boosting که از روش‌های یادگیری جمعی هستند به پیش‌بینی جریان روزانه رودخانه سیماو در حوضه سوسرلک ترکیه پرداختند. طبق نتایج به دست آمده، مدل‌های یادگیری جمعی در برآورد داده‌های روزانه جریان موفق هستند و می‌توان داده‌های جریان را با استفاده از داده‌های جریان رودخانه‌های فرعی تخمین زد

زیرا عملکرد مدل بهبود می‌یابد. همچنین، محققان برای پیش‌بینی و جلوگیری از خسارات ناشی از سیلاب در آنکارا از مدل‌های مختلفی مانند RF، K-NN، Xgboost، BTress و SVM استفاده کرده‌اند. نتایج حاصل از معیارهای ارزیابی، نشان‌دهنده آن است که K-NN دارای کمترین میزان خطا و بهترین عملکرد در پیش‌بینی سیلاب است (Katipoğlu and Sarigol, 2023).

Terzi et al. (2023) با تمرکز بر سه مدل RFR (رگرسیون جنگل تصادفی) ، Xgboost و LSTM (حافظه بلندمدت و کوتاه‌مدت) به تخمین جریان روزانه رودخانه گوکسو با استفاده از دوره آماری ۲۰ ساله (۱۹۹۰ تا ۲۰۱۰) پرداختند. جهت تنظیم فراپارامترهای مدل از روش جستجوی شبکه (Grid Search) استفاده شد. نتایج معیارهای ارزیابی R^2، RMSE و MAE نشان داد که مدل Xgboost عملکرد بهتری نسبت به مدل‌های RFR و LSTM دارد.

Dariane and Borhan (2024) به مقایسه روش‌های کلاسیک و یادگیری ماشین در تخمین داده‌های مفقود در سه حوضه کوهستانی شمال ایران (طالقان، کرج و لتیان) با استفاده از داده‌های ۲۶ ساله (۱۳۷۰ تا ۱۳۹۶) پرداختند. روش‌های یادگیری ماشین شامل رگرسیون خطی ساده (LR) و رگرسیون خطی چندگانه (MLR)، شبکه عصبی مصنوعی (ANN) با پنج ساختار مختلف، رگرسیون بردار پشتیبان (SVR)، درخت تصمیم M5 و سیستم استنتاج عصبی فازی تطبیقی (ANFIS) می‌باشد. به طور کلی، مشاهده شد که روش‌های مبتنی بر یادگیری ماشین خروجی بهتری دارند.

همچنین، سیدیان و همکاران (۱۳۹۳) جریان رودخانه گرگانرود را در دو مقیاس زمانی هفتگی و ماهانه در سه ایستگاه هیدرومتری حاجی‌قوشان، قره‌شور و تمر برای ۹ سال آماری شبیه‌سازی کردند. نتایج نشان داد که مدل SVM نسبت به مدل سری زمانی مقدار RMSE را ۳۵ درصد کاهش داده است. آهنی و شوریان (۱۳۹۶) عملکرد مدل‌های داده مبنا جهت پیش‌بینی جریان ماهانه رودخانه در ایستگاه سراب هنده در حوضه آبریز دریاچه نمک را بررسی کردند. نتایج حاکی از برتری مدل K-NN بین مدل‌های رگرسیون خطی چندگانه، شبکه عصبی و رگرسیون خطی چندگانه بود.

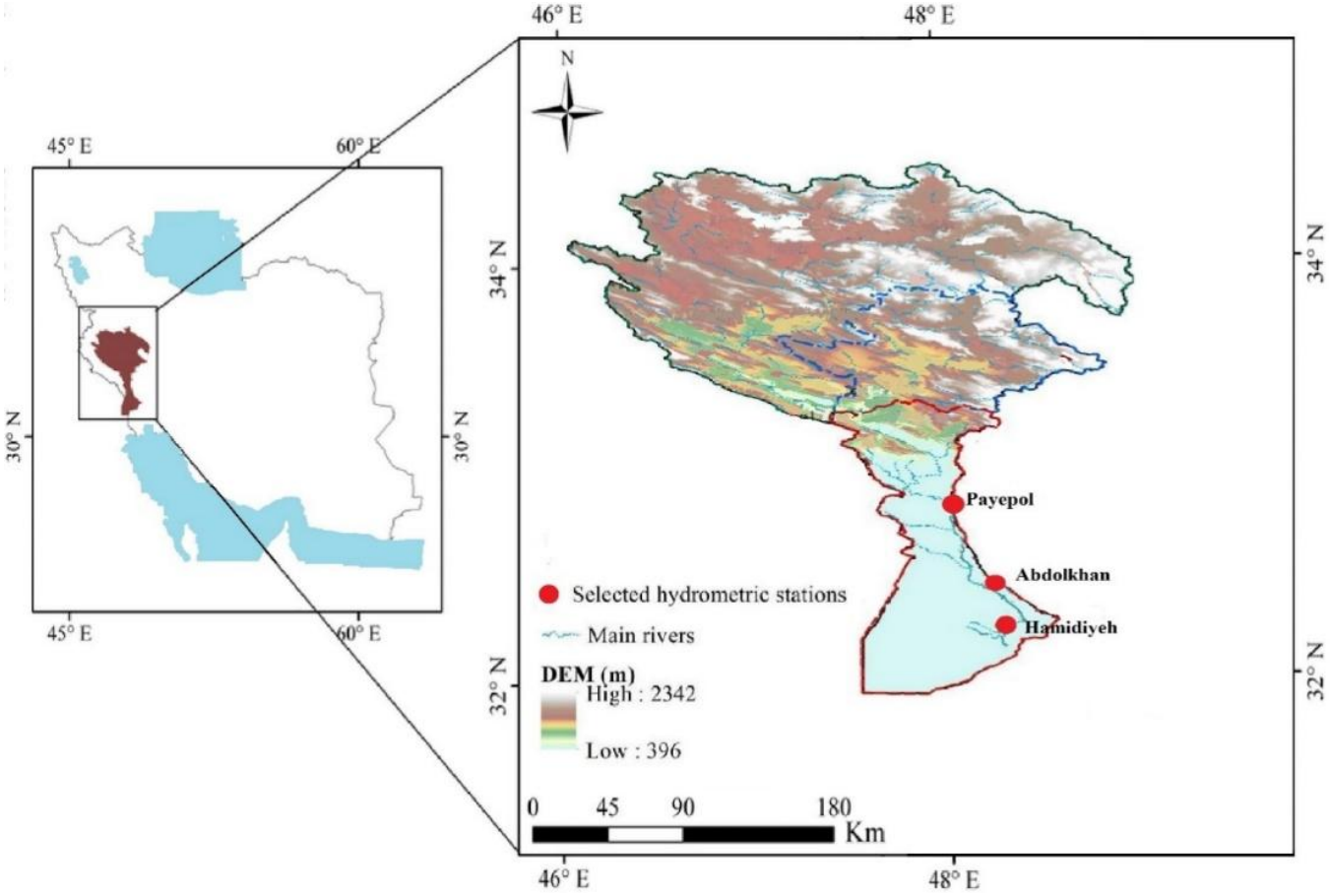
دارابی و همکاران (۱۴۰۳) در پژوهش خود بیان داشتند که استفاده از مدل‌های یادگیری ماشین سبب صرفه‌جویی در زمان و هزینه برای پروژه‌های پیش‌بینی جریان رودخانه و مدیریت سیلاب در حوضه‌های آبخیز است.

بررسی تحقیقات گذشته نشان می‌دهد که روش‌های مبتنی بر یادگیری ماشین برای بازسازی مقادیر مفقود مناسب‌ترین هستند و معمولاً منجر به بهبود قابل توجهی در دقت پیش‌بینی نسبت به روش‌های مبتنی بر رویکردهای آماری می‌شوند (Krysanova & White, 2015; Minns & Hall, 1996; Varga et al., 2016). علاوه بر این، انتخاب مدل بهینه و تنظیم مؤثر فراپارامترها نقش مهمی در بهبود عملکرد این مدل‌ها دارد. تاکنون تحقیقات محدودی در حوزه مهندسی آب، از روش Optuna برای بهینه‌سازی فراپارامترهای مدل‌های یادگیری ماشین استفاده شده است. در این پژوهش، از روش Optuna به منظور تنظیم فراپارامترهای مدل‌های یادگیری ماشین در تخمین داده‌های مفقود استفاده شده است. همچنین، با توجه به اینکه حوضه کرخه از مهم‌ترین حوضه‌های آبریز کشور ایران و مهم‌ترین منبع تأمین‌کننده آب بخش‌های مختلف و نواحی مجاور خود از لحاظ کشاورزی و شرب می‌باشد و کاهش جریان رودخانه‌های حوضه آبریز مذکور سبب مشکلات زیادی در حوضه آبریز کرخه شده است، بنابراین، اهمیت شبیه‌سازی و تخمین داده‌های جریان رودخانه مذکور، بیش از پیش ضروری است. در پژوهش حاضر، با استفاده از طیف وسیعی از مدل‌های یادگیری ماشین و مدل‌های یادگیری جمعی مبتنی بر یادگیری ماشین، به بررسی توانمندی ۹ مدل متفاوت جهت ترمیم و تخمین داده‌های مفقود در دو ایستگاه هیدرومتری پرداخته می‌شود. هدف اصلی این مطالعه، معرفی روش کارآمد جهت تخمین داده‌های مفقود حوضه رودخانه کرخه و بررسی کارایی مدل‌ها در شرایط چالش وجود داده‌های محدود و همچنین تأثیر فاصله مکانی ایستگاه مبنا با ایستگاه دارای داده مفقود، بر دقت مدل‌ها می‌باشد. در ادامه، فراپارامترهای بهینه مدل‌ها جهت تخمین جریان روزانه، به روش Optuna تنظیم می‌شود و بهترین ساختار مدل‌ها تعیین می‌گردد. سپس، عملکرد آنها با استفاده از شاخص‌های آماری و بصری مورد ارزیابی قرار می‌گیرد.

مواد و روش‌ها

منطقه مورد مطالعه

حوضه آبریز کرخه به لحاظ تقسیم‌بندی هیدرولوژیک جزئی از حوضه آبریز خلیج فارس بوده و در غرب ایران و در مناطق جنوبی و میانی رشته کوه زاگرس و مساحتی بالغ بر ۵۰۷۶۸ کیلومتر مربع قرار دارد. این حوضه از نظر مختصات جغرافیایی، بین ۴۶ درجه و ۷ دقیقه تا ۴۹ درجه و ۱۱ دقیقه طول شرقی و ۳۲ درجه و ۲۵ دقیقه تا ۳۴ درجه و ۵۶ دقیقه عرض شمالی قرار گرفته است. تصویر موقعیت مکانی ایستگاه‌های هیدرومتری مورد مطالعه در شکل ۱ آورده شده است. در جدول ۱، مختصات جغرافیایی ایستگاه‌های هیدرومتری مورد مطالعه نمایش داده شده است. در جدول ۲ نیز مشخصات آماری ایستگاه‌ها ارائه شده است. تعداد روزهای دارای داده مفقود در ایستگاه‌های هیدرومتری عبدالخان و پای پل به ترتیب ۳۴۷۵ و ۷۹۰ روز است.



شکل(۱): موقعیت جغرافیایی ایستگاه‌های هیدرومتری مورد مطالعه.

جدول(۱): مشخصات ایستگاه‌های هیدرومتری حمیدیه، عبدالخان و پای پل.

UTM_Y	UTM_X	Longitude	Latitude	Code	Station
۳۴۸۷۸۷۰/۷۸۹	۲۵۶۲۲۰/۹۵	۴۸-۲۶-۰۰	۳۱-۳۰-۰۰	۲۱-۱۹۹	Hamidiyeh
۳۵۸۸۳۹۰/۲۷۱	۲۳۰۳۶۱/۸۶۸	۴۸-۰۸-۰۰	۳۲-۲۴-۰۰	۲۱-۱۹۱	Paye-pol
۳۵۲۴۹۴۶/۱۲۱	۲۵۲۳۵۷/۳۳۶	۴۸-۲۳-۰۰	۳۱-۵۰-۰۰	۲۱-۱۹۳	Abdolkhan

جدول (۲): مشخصات آماری داده‌های دبی جریان روزانه رودخانه کرخه در ایستگاه‌های مورد بررسی.

Skew	Standard deviation (m^3/s)	Average discharge (m^3/s)	Min (m^3/s)	Max (m^3/s)	Number of data	Station
۳/۲۱	۲۰۰/۹۳	۱۵۸/۰۹	۱	۲۳۵۷	۱۴۹۷۵	Hamidiyeh
۴/۷۰	۲۰۶/۴۸	۱۶۸/۷۱	۱	۳۷۱۹	۱۱۵۰۰	Abdolkhan
۵/۲۰	۲۳۳/۸۰	۱۸۳/۷۰	۱.۵	۵۲۰۳	۱۴۱۸۵	Paye-pol

جهت مدل‌سازی، داده‌ها به دو دسته، داده‌های آموزشی و داده‌های آزمایشی تقسیم شدند. در این پژوهش، ۷۰٪ از کل داده‌ها به آموزش مدل تعلق گرفت و ۳۰٪ باقیمانده به عنوان داده‌های آزمون به مدل معرفی گردید. انتخاب داده‌های آموزشی و آزمایشی توسط برنامه و کتابخانه sklearn انجام شد. برای تنظیم فراپارامترها در الگوریتم‌های یادگیری ماشین از روش Optuna استفاده شد که شامل جستجو در محدوده‌ای از مقادیر فراپارامتر برای یافتن ترکیب بهینه آن است که بهترین عملکرد را در یک مجموعه داده معین ایجاد می‌کند (Akiba et al., 2019). در این پژوهش، از زبان برنامه نویسی پایتون، که یک زبان قدرتمند با طیف گسترده‌ای از کتابخانه‌ها و ابزارها محسوب می‌شود، استفاده شده است.

الگوریتم تقویت گرادیان شدید

الگوریتم Xgboost (یا XGB) یک رویکرد یادگیری جمعی است که از مجموعه یادگیرهای ضعیف، یک یادگیرنده قوی می‌سازد (Chemura et al., 2020). مدل Xgboostیک گام بهبودیافته از روش جزءبندی مبتنی بر درخت بازگشتی ارتقای گرادیان است که توسط (2001) Friedmanمعرفی گردید و پس از توسعه توسط (2016) Guestrin and Chenبه عنوان یکی از الگوریتم‌های بسیار کارآمد یادگیری ماشین برای زمینه‌های مختلف شناخته شده (Chemura et al., 2020) و در مدل‌های رگرسیونی و طبقه‌بندی مورد استفاده گسترده قرار گرفته است .

(Khoramipoor et al., 2024)معادلات این مدل به صورت روابط (۱) تا (۳) می‌باشد (Fan et al., 2018).

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \quad (۱)$$

که در آن، $f_t(x_i)$ یادگیرنده در مرحله t، $f_i^{(t)}$ و $f_i^{(t-1)}$ پیش‌بینی‌های مراحل t و (t-1) هستند و x_i متغیر ورودی است.

$$Obj^{(t)} = \sum_{k=1}^{n} l(\overline{y_i}, y_i) + \sum_{k=1}^{t} \Omega(f_i) \quad (۲)$$

که در آن، l تابع از دست دادن، n تعداد مشاهدات استفاده شده و Ω عبارت منظم‌سازی است که به جلوگیری از Overfitting و پیچیدگی مدل کمک می‌کند و به صورت رابطه (۳) تعریف می‌شود:

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \quad (۳)$$

که در آن ω بردار امتیاز در برگ‌ها، λ پارامتر تنظیم کننده و γ حداقل میزان کاهش خطا برای تقسیم بیشتر گره برگ است.

الگوریتم Catboost

الگوریتم Catboost روشی نوین است که از تقویت گرادیان مبتنی بر درخت برای حل مسائل رگرسیون و طبقه‌بندی استفاده می‌کند. این روش تقویت گرادیان، الگوریتمی را آموزش می‌دهد که در هر تکرار با کمینه‌سازی گرادیان، بهینه‌سازی تابع زیان را انجام می‌دهد. اما این فرایند ممکن است سبب برازش بیش از حد مدل شود. به منظور حل این مشکل، Catboost به کمک روش تقویت

سفارشی (Ordered boosting) برای تخمین گرادیان، بایاس (اُریبی) مدل را کاهش می‌دهد (Zhang et al., 2020). همچنین، به منظور کنترل برازش بیش از حد مدل در ساختار درختی، با ایجاد جایگشت‌های تصادفی، تخمین مقادیر درختان تصمیم را انجام می‌دهد (Jabeur et al., 2021).

الگوریتم جنگل تصادفی

الگوریتم جنگل تصادفی (Random Forest, RF) را نخستین بار Breiman در سال ۲۰۰۱ ایجاد نمود. جنگل تصادفی، چندین درخت تصمیم را می‌سازد و آنها را با یکدیگر ادغام می‌کند تا پیش‌بینی‌های صحیح‌تر به دست آیند. مدل پیش‌بینی کننده جنگل تصادفی بر اساس میانگین‌گیری از نتایج حاصل از تمامی درخت‌های تصمیم مربوطه استوار است و از دو عامل میانگین کاهشی دقت و میانگین کاهشی جینی برای تعیین اولویت تأثیر هر یک از عوامل مؤثر استفاده می‌شود (Yohaness, 1999).

الگوریتم درختان اضافه

الگوریتم درختان اضافه (Extra Trees, ET) یک روش یادگیری جمعی است، که از چندین درخت تصمیم برای پیش‌بینی استفاده می‌کند و هنگام انتخاب بخش‌های هر گره، به طور تصادفی زیرمجموعه‌ای از ویژگی‌ها را از کل نمونه اصلی انتخاب می‌کند. این الگوریتم به طور تصادفی نقطه تقسیم را انتخاب می‌کند ولی نقطه بهینه را انتخاب نمی‌کند. به این دلایل، درختان بسیار تصادفی یا به طور خلاصه درختان اضافی نامیده شده است (Geurts et al., 2006).

الگوریتم شبکه‌های عصبی

شبکه پرسپترون چندلایه (MLP) نوعی از شبکه‌های عصبی مصنوعی (Artificial Neural Networks, ANNs) است که بر مبنای یک واحد محاسباتی به نام پرسپترون ساخته می‌شود. یک پرسپترون، برداری از ورودی‌ها را دریافت می‌کند. اگر مقدار حاصل از یک مقدار آستانه بیشتر بود، خروجی پرسپترون برابر با ۱ و در غیر این صورت، معادل ۱- خواهد بود. این شبکه شامل یک لایه ورودی، یک یا چند لایه پنهان و یک لایه خروجی است. برای آموزش این شبکه، از الگوریتم پس‌انتشار خطا (BP) استفاده می‌شود. طی آموزش شبکه MLP به کمک الگوریتم یادگیری BP، ابتدا محاسبات از ورودی شبکه به سوی خروجی شبکه انجام می‌شود و سپس، مقادیر خطای محاسبه شده به لایه‌های قبل انتشار می‌یابد و وزن‌ها تصحیح می‌شوند. ابتدا، خروجی به صورت لایه به لایه محاسبه می‌شود و خروجی هر لایه، ورودی لایه بعدی خواهد بود (غفاری و وفاخواه، ۱۳۹۲).

الگوریتم k نزدیک‌ترین همسایه (KNN)

این الگوریتم با شناسایی k نزدیک‌ترین همسایه یک نقطه داده جدید بر اساس متریک فاصله (مثلاً فاصله اقلیدسی) کار می‌کند (Sumayli, 2023). سپس به نقطه داده جدید برچسب کلاسی اختصاص داده می‌شود که در میان k نزدیکترین همسایه آن بیشترین فراوانی را دارد (Xie et al., 2024). مدل رگرسیون نزدیکترین همسایه با تعیین فاصله بین یک مشاهده جدید و همه مشاهدات موجود در داده‌های آموزشی عمل می‌کند (محمدی و همکاران، ۱۴۰۰).

الگوریتم درخت تصمیم

الگوریتم CART (Classification And Regression Trees) به عنوان یکی از متداول‌ترین روش‌های درخت تصمیم‌گیری توسط Breiman و همکاران (2017) به وجود آمد، که برخلاف مدل‌های شبکه عصبی، به تولید قانون می‌پردازد. CART یک روش سلسه مراتبی یا چند مرحله‌ای است که در آن به صورت بازگشتی مجموعه داده‌ها به روش دودویی به تقسیمات فرعی و کوچک‌تر تقسیم‌بندی می‌شوند. تقسیم‌بندی‌ها تا زمانی که تقسیمات فرعی نهایی نتوانند بیشتر از آن تجزیه شوند ادامه می‌یابد. سپس، عملیات هرس کردن درخت انجام می‌شود (یوسفی و همکاران، ۱۴۰۱).

رگرسیون بردار پشتیبان (SVR)

این روش توسط Boser و همکاران در سال ۱۹۹۲ بر پایه تئوری یادگیری آماری با هدف حداقل‌سازی خطای ساختاری و تولید یک جواب بهینه معرفی شده است. ماشین‌های بردار پشتیبان برای حل مسائل غیرخطی، ابعاد مسئله را از طریق توابع کرنل تغییر می‌دهند. انتخاب کرنل برای ماشین بردار پشتیبان به حجم داده‌های آموزشی و ابعاد بردار ویژگی بستگی دارد.

مدل درختی M5P

مدل درخت M5P توسط Quinlan جهت پیش‌بینی داده‌های پیوسته ارائه شده است (Quinlan, 1992). در مرحله اول، برای ساخت درخت تصمیم از یک الگوریتم استقرایی یا معیار انشعاب استفاده می‌شود. همچنین، مقدار کاهش مورد انتظار در این خطا را تحت عنوان نتیجه آزمون هر صفت در گره محاسبه می‌کند. مرحله دوم، ساخت مدل درختی، هرس کردن درخت توسعه یافته و جایگزین کردن درختان فرعی به جای درخت اصلی، با توابع رگرسیونی می‌باشد. پس از بیشینه کردن کل انشعابات ممکن، M5P صفتی که کاهش مورد انتظار را حداکثر کند، انتخاب می‌کند. این تقسیم باعث ایجاد ساختار شبه‌درختی بزرگی می‌شود که بیش‌برازش مدل را در پی دارد. برای غلبه بر این مسئله، از مرحله دوم هرس کردن که به صورت جایگزین کردن یک درخت فرعی به جای یک برگ است، استفاده می‌شود (میرنورالهی، ۱۴۰۱).

روش‌های ارزیابی مدل

عملکرد مدل‌های مورد استفاده در این پژوهش با استفاده از معیارهای آماری ریشه میانگین مربع خطاها (RMSE)، مجذور میانگین مربعات خطای نسبی (RRMSE)، میانگین خطای مطلق (MAE) و ضریب تعیین (R^2) ارزیابی شده است. معادلات معیارهای ارزیابی مذکور به شرح روابط (۴) تا (۷) می‌باشند.

$$MAE=\frac{\sum_{i=1}^{N}|y_i-x_i|}{N} \quad \text{Mean Absolute Error} \quad (۴)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2} \quad \text{Root Mean Squared Error} \quad (۵)$$

$$RRMSE=\sqrt{\frac{\sum_{i=1}^{N}(x_i-y_i)^2}{N*Std(y_i)^2}} \quad \text{Relative Root Mean Squared Error} \quad (۶)$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-x_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \quad \text{Coefficient of determination} \quad (۷)$$

در این روابط y_i مقدار واقعی جریان رودخانه برای نمونه i، x_i مقدار پیش‌بینی شده جریان رودخانه برای نمونه i، $\overline{y}$ مقدار متوسط جریان واقعی، $\overline{x}$ مقدار متوسط جریان پیش‌بینی شده و N تعداد نمونه‌ها است.

دیاگرام تیلور

نمودار بسط تیلور، نمودار ریاضی است که برای مقایسه گرافیکی عملکرد مدل‌های مختلف در پیش‌بینی متغیرهای پیوسته استفاده می‌شود. فاصله بین هر مدل و نقطه مرجع (داده‌های مشاهده شده) معیاری است برای اینکه هر مدل تا چه حد داده‌ها را واقع‌بینانه شبیه‌سازی نموده است. این نمودار، خلاصه‌ای واضح و سریع از میزان مطابقت الگو را نشان می‌دهد و به سادگی می‌توان میزان دقت مدل‌های شبیه‌سازی سیستم را مشخص کرد (Taylor, 2001).

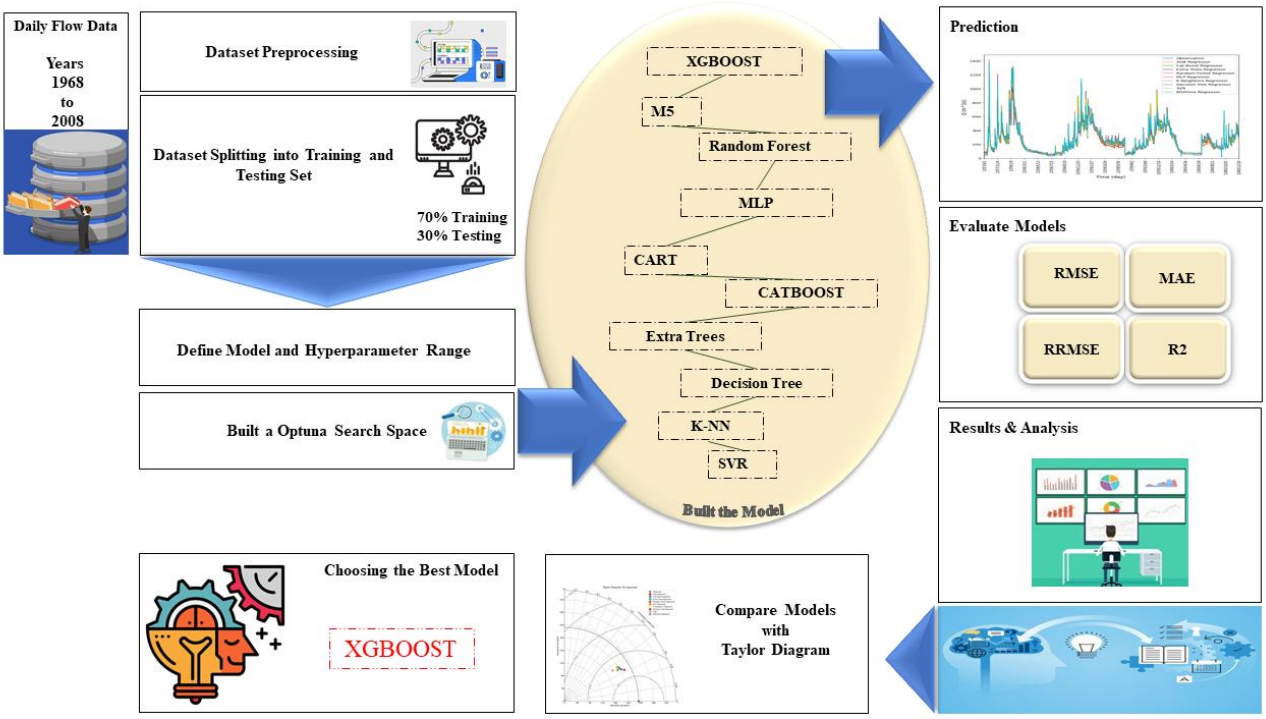
روش بهینه‌سازی Optuna

بهینه‌سازی فراپارامترها یکی از مراحل کلیدی در توسعه مدل‌های یادگیری ماشین است که نقش اساسی در دستیابی به عملکرد بهینه ایفا می‌کند. Optuna به‌عنوان

یک کتابخانه در زبان برنامه‌نویسی پایتون، ابزاری کارآمد و کاربر پسند برای بهینه‌سازی فراپارامترها ارائه می‌دهد.

در این پژوهش، فضای جستجو برای پارامترها تعریف و بهترین مجموعه پارامترها برای حداقل کردن معیار ارزیابی انتخاب شد. روش بهینه‌سازی Optuna به دلیل کارایی بالا و توانایی کاهش بار محاسباتی نسبت به روش‌های سنتی مانند جستجوی شبکه‌ای انتخاب گردید.

فلوچارت مراحل اجرای پژوهش در شکل ۲ ارائه گشته است.



شکل(۲): فلوچارت مراحل انجام پژوهش.

نتایج و بحث

در این بخش، دبی جریان رودخانه کرخه در سه ایستگاه هیدرومتری پای پل، عبدالخان و حمیدیه با استفاده از تکنیک‌ها و عملیات داده‌کاوی ۹ مدل از الگوریتم‌های یادگیری جمعی و ماشینی از جمله Catboost، Xgboost، Random forest، Extra trees، درخت تصمیم مدل M5، درخت تصمیم مدل Cart، K-NN، MLP و SVR مورد کاوش قرار گرفت تا الگوهای مورد نظر کشف گردند. جهت تخمین داده‌های مفقود ایستگاه‌های هیدرومتری عبدالخان و پای پل، از داده‌های مشاهداتی دبی جریان ایستگاه همسایه آنها، یعنی ایستگاه هیدرومتری حمیدیه، به عنوان یکی از ویژگی‌های داده‌های آموزشی استفاده شده است. ایستگاه هیدرومتری حمیدیه به عنوان ایستگاه مبنا بر اساس دو معیار انتخاب شد: ۱) فاصله کم و ۲) داشتن اطلاعات در خلأهای ایستگاه‌های هیدرومتری عبدالخان و پای پل. تنظیم دقیق فراپارامترها می‌تواند عملکرد تخمین مدل‌ها را به مقدار قابل توجهی بهبود بخشد. در تمام مدل‌ها، از روش Optuna برای یافتن

مقادیر پارامترها استفاده شد (Akiba et al., 2019). پس از انتخاب ویژگی‌ها و تنظیم فراپارامترها، مقایسه عملکرد مدل‌ها جهت تخمین داده‌های مفقود جریان رودخانه ایستگاه‌های عبدالخان و پای پل صورت گرفت.

جداول ۳ و ۴ مقادیر دقت به دست آمده از ارزیابی الگوریتم‌های استفاده شده با معیارهای مختلف صحت‌سنجی، برای ایستگاه‌های هیدرومتری عبدالخان و پای پل در دوره آموزش و آزمون را نشان می‌دهند. بر اساس نتایج جداول مذکور، در دوره آموزش برای ایستگاه عبدالخان، مقدار معیار MAE برای الگوریتم‌های Catboost، Xgboost، Random forest، Extra trees، درخت تصمیم مدل M5، درخت تصمیم مدل Cart، K-NN، MLP و SVR به‌ترتیب برابر ۷/۹۳۶، ۱۱/۰۹، ۱۰/۲۳۳، ۱۲/۹۷۷، ۱۹/۳۱۹، ۱۹/۲۶۵، ۲۰/۶۷۷، ۲۵/۸۲۸ و ۲۵/۳۸۰ است. معیار RMSE برای الگوریتم‌های ذکر شده به‌ترتیب ۱۲/۹۸، ۱۸/۲۱، ۳۱/۵۶، ۳۱/۸۵، ۴۹/۹۵، ۵۴/۷۳، ۵۸/۹۳، ۶۵/۵۵ و ۷۴/۷۷ محاسبه گردید. معیار RRME به‌ترتیب الگوریتم‌های مذکور برابر ۰/۰۶، ۰/۰۸، ۰/۱۵، ۰/۱۵، ۰/۲۴، ۰/۲۶، ۰/۲۸، ۰/۳۲ و ۰/۳۶ محاسبه شده است. مقادیر R^2 نیز برابر ۰/۹۹، ۰/۹۹، ۰/۹۷، ۰/۹۷، ۰/۹۳، ۰/۹۲، ۰/۹۱، ۰/۸۹ و ۰/۸۶ می‌باشد. به همین ترتیب در دوره آموزش، برای ایستگاه هیدرومتری پای پل، مقدار معیار MAE برای الگوریتم‌های Xgboost، Catboost، Extra trees، Random forest، M5، K-NN، Decision tree، MLP و SVR به‌ترتیب برابر با ۱۷/۶۴، ۲۵/۰۴، ۲۸/۶۳، ۲۵/۵۶، ۳۸/۷۴، ۴۰/۱۳، ۴۰/۳۹، ۴۲/۴۰ و ۳۹/۲۳ بود. معیار RMSE برای الگوریتم‌های مذکور به‌ترتیب ۳۱/۶۷، ۶۱/۱۱، ۸۸/۳۰، ۹۰/۸۶، ۱۲۳/۷۵، ۱۲۵/۹۶، ۱۲۶/۱۹، ۱۲۶/۲۷ و ۱۳۰/۶۷ به‌دست آمد. مقدار R^2 نیز برای الگوریتم‌های مذکور به‌ترتیب برابر ۰/۹۸، ۰/۹۲، ۰/۸۵، ۰/۸۴، ۰/۷۲، ۰/۷۱، ۰/۷۰، ۰/۷۰ و ۰/۶۸ محاسبه شد. بنابراین، با توجه به معیارهای ارزیابی در دوره آموزش، دقت الگوریتم‌های مورد استفاده در هر دو ایستگاه هیدرومتری مورد مطالعه قابل قبول است.

جهت انتخاب بهترین الگوریتم، از داده‌های جدید برای دوره آزمون استفاده شد. همانطور که مشاهده می‌شود، بر اساس نتایج دوره آزمون در جدول ۳ برای ایستگاه عبدالخان، الگوریتم Xgboost از دقت بیشتری نسبت به سایر مدل‌ها برخوردار است. الگوریتم Xgboost با داشتن کمترین میانگین خطای مطلق، ریشه میانگین مربع خطاها و مجذور میانگین مربعات خطای نسبی به ترتیب برابر ۱۸/۷۶، ۴۳/۷۵، ۰/۲۰ و بیشترین ضریب تعیین ۰/۹۵ در بین الگوریتم‌های داده‌کاوی استفاده شده در این پژوهش دارای عملکرد بهتر می‌باشد. با توجه به اینکه مقادیر پیش‌فرض داده شده در مدل‌های یادگیری ماشین بهترین عملکرد را تضمین نمی‌کند (Schratz et al., 2019)، بنابراین تنظیم دقیق فراپارامترها می‌تواند عملکرد مدل را به مقدار قابل توجهی بهبود ببخشد. در تمام مدل‌ها، از روش Optuna برای یافتن بهترین مقادیر پارامترها استفاده شد (Akiba et al., 2019). جهت کاهش فضای جستجو، فراپارامترهای مهم مدل‌ها بهینه‌سازی شدند.

در ایستگاه پای پل نیز نتایج حاکی از عملکرد بهتر الگوریتم Xgboost با مقادیر میانگین خطای مطلق، ریشه میانگین مربع خطاها و مجذور میانگین مربعات خطای نسبی به ترتیب برابر ۳۶/۴۵، ۱۰۸/۸۷، ۰/۴۶ و بیشترین ضریب تبیین برابر ۰/۷۸ به دست آمده است. بررسی معیارهای ارزیابی مختلف حاکی از عملکرد بهتر مدل Xgboost در هر دو ایستگاه هیدرومتری عبدالخان و پای پل می‌باشد. پس مدل Xgboost در حوضه آبخیز کرخه در هر دو دوره آموزشی و آزمون به بهترین شکل شبیه‌سازی می‌کند. در ایستگاه پای پل، عملکرد مدل‌های K-NN، Decision Tree و SVR در دوره صحت‌سنجی نسبت به دوره آموزش بهبود داشته است. برای مثال، مقادیر RMSE مدل‌های مذکور به ترتیب از ۱۲۵/۹۶، ۱۲۶/۱۹ و ۱۳۰/۶۷ متر مکعب در ثانیه برای آموزش به مقدار ۱۲۴/۷۶، ۱۲۴/۶۱ و ۱۲۷/۴۱ متر مکعب در ثانیه در دوره صحت‌سنجی کاهش یافته است. این نتایج با نتایج Samadi et al. (2019) و Mohammadi et al. (2021) همخوانی دارد، که در مطالعات آن‌ها نیز عملکرد مدل در مرحله صحت‌سنجی بهتر از مرحله آموزش بوده است. با این وجود، عملکرد کلی مدل‌های مذکور در ایستگاه پای پل، با وجود بهبود عملکرد در مرحله آزمون، متوسط ارزیابی شد و فقط مدل‌های Xgboost و Catboost دارای عملکرد خوبی در ایستگاه مذکور بودند. همچنین، مدل SVR در ایستگاه عبدالخان به ترتیب با مقادیر R^2، RRMSE، RMSE و MAE برابر ۰/۸۴، ۰/۳۹، ۸۲/۹۳ و ۲۷/۰۴ و در ایستگاه

پای پل به ترتیب با ۰/۷۰، ۰/۵۴، ۱۲۷/۴۱ و ۳۸/۵۲ ضعیف‌ترین عملکرد را در بین مدل‌های Catboost، Xgboost، Random forest، Extra trees، درخت تصمیم مدل M5، درخت تصمیم مدل Cart، K-NN و MLP دارد.

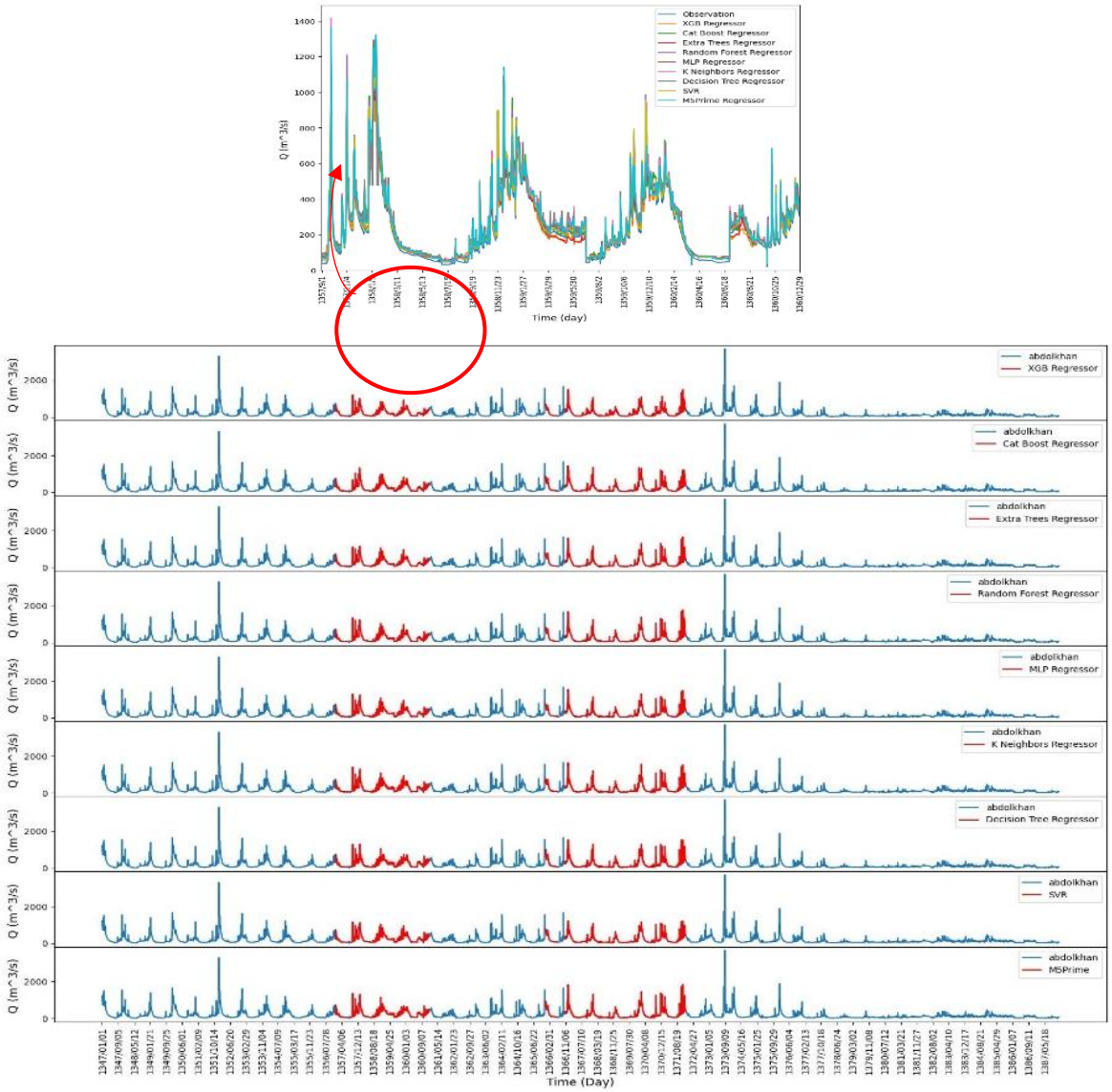
نتایج مقایسه عملکرد مدل‌ها در هر دو ایستگاه عبدالخان و پای پل نشان می‌دهد با اینکه ایستگاه عبدالخان دارای ۱۵ سال داده مفقود از ۴۰ سال دوره مطالعه در این پژوهش می‌باشد، اما به دلیل فاصله کم تا ایستگاه همسایه (۲۲/۴ کیلومتر تا ایستگاه هیدرومتری حمیدیه)، همه مدل‌ها توانسته‌اند به خوبی داده‌های مفقود را تخمین بزنند؛ به طوری که مدل Xgboost، با مقدار ۰/۹۵ بیشترین و مدل SVR با مقدار ۰/۸۴ کمترین مقدار R^2 را در بین ۹ مدل مورد بررسی داشته‌اند. همچنین، تمام مدل‌ها دارای مقادیر RRMSE کمتر از ۰/۵ هستند که نشان‌دهنده عملکرد خوب آنها می‌باشد. اما در ایستگاه پای پل، تنها با داشتن دو سال داده مفقود از ۴۰ سال دوره مطالعه، با توجه به نتایج RRMSE، فقط دو مدل Xgboost و Catboost توانستند مقادیر کمتر از ۰/۵ به‌دست آورند و عملکرد خوبی داشته باشند، که این نشان می‌دهد تنها الگوریتم‌های مذکور در بین الگوریتم‌های بررسی شده قادر به غلبه بر چالش‌های مکانی بودند. که این امر می‌تواند به علت ساختار الگوریتم‌های Boosting باشد که به صورت متوالی آموزش می‌بینند و مدل‌های بعدی خطای مدل‌های قبلی را اصلاح می‌کنند. این آموزش متوالی سبب خطای کمتر و دقت بیشتر شده است. مقادیر RRMSE برای دو مدل مذکور به ترتیب برابر ۰/۴۶ و ۰/۴۹ به‌دست آمد. بقیه مدل‌ها مقادیری بین ۰/۵ الی ۰/۵۴ به‌دست آوردند که نشان‌دهنده عملکرد متوسط آنها است. به نظر می‌رسد که عملکرد متوسط اغلب مدل‌ها در ایستگاه پای پل به علت فاصله زیاد تا ایستگاه مبنا (ایستگاه هیدرومتری حمیدیه) می‌باشد. بنابراین، مقایسه عملکرد مدل‌ها در دو ایستگاه با شرایط متفاوت فاصله مکانی تا ایستگاه مبنا، نشان می‌دهد، الگوریتم‌های Boosting توانسته‌اند تأثیر فاصله زیاد از ایستگاه مبنا را نادیده گرفته و عملکرد خوبی داشته باشند. بررسی اینکه فاصله بین ایستگاه‌ها چگونه بر دقت مدل‌ها اثر می‌گذارد، می‌تواند به ما در طراحی بهتر شبکه‌های پایش هیدرومتری کمک کند. این کار به مدیران این امکان را می‌دهد که شبکه‌هایی کارآمدتر بسازند و تصمیم‌های هوشمندانه‌تری درباره مکان قرارگیری ایستگاه‌های جدید بگیرند. شکل‌های ۳ و ۴ نمودارهای سری زمانی مقادیر مشاهداتی جریان روزانه رودخانه کرخه در دو ایستگاه هیدرومتری عبدالخان و پای پل در دوره آزمون را نشان می‌دهند. داده‌های مفقود تخمین زده شده با رنگ قرمز و داده‌های مشاهداتی با رنگ آبی نمایش داده شده‌اند. بخشی از نمودار مربوط به تخمین داده‌های مفقود جریان رودخانه با استفاده از الگوریتم Xgboost بزرگ‌نمایی شده است تا جزئیات دقیق‌تر و وضوح بهتری از عملکرد مدل، نمایش داده شود. این بخش از نمودار به وضوح نشان می‌دهد که مدل Xgboost در پیش‌بینی داده‌های مفقود توانسته است به‌خوبی با مشاهدات واقعی تطابق داشته باشد و دقت بالای خود را در تخمین جریان‌ها نشان می‌دهد.

جدول (۳): مقایسه مدل‌های مورد بررسی در ایستگاه هیدرومتری عبدالخان با استفاده از معیارهای ارزیابی مختلف.

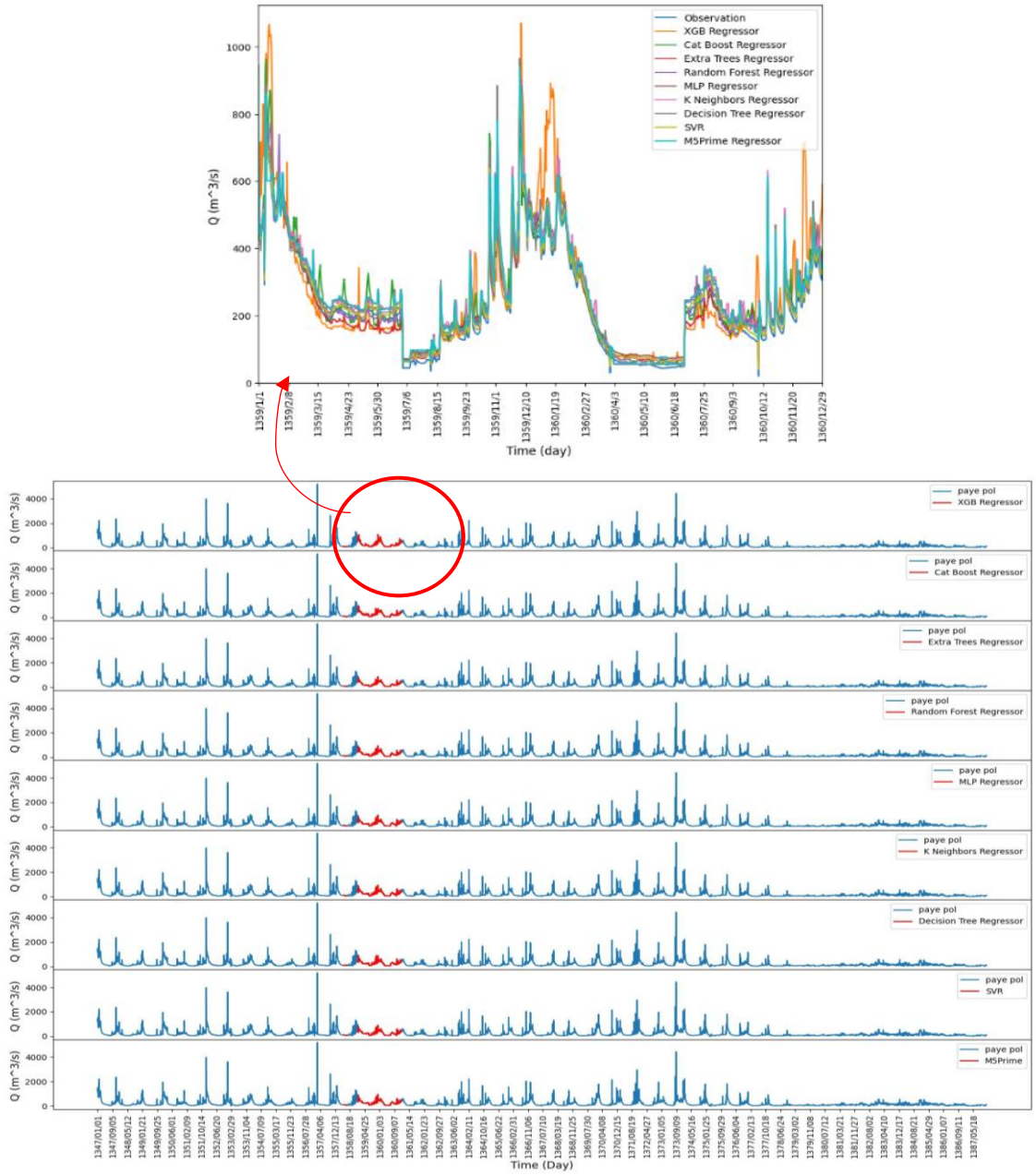
	Model	R^2	RRMSE	RMSE	MAE
Training	Cat Boost	۰/۹۹۵۴۳۹	۰/۰۶۳۵۹۱	۱۲/۹۸۵۳۲	۷/۹۳۶۱۳۹
	XGB	۰/۹۹۱۴۱۹	۰/۰۸۹۲	۱۸/۲۱۴۶۳	۱۱/۰۶۰۱۶
	Random Forest	۰/۹۷۵۱۶۵	۰/۱۵۴۵۸۶	۳۱/۵۶۶۵۵	۱۰/۲۳۳۶۷
	Extra Trees	۰/۹۷۲۹۵۲	۰/۱۵۶۰۰۸	۳۱/۸۵۶۸۸	۱۲/۹۷۷۷۲
	M5Prime	۰/۹۳۹۹۳۳	۰/۲۴۴۶۵۷	۴۹/۹۵۹۱۹	۱۹/۳۱۹۹۱
	Decision Tree	۰/۹۲۸۱۳۱	۰/۲۶۸۰۶۲	۵۴/۷۳۸۴۴	۱۹/۲۶۵۵۷
	K Neighbors	۰/۹۱۶۶۳۸	۰/۲۸۸۶۰۲	۵۸/۹۳۲۶۲	۲۰/۶۷۷۹
	MLP	۰/۸۹۶۷۸۱	۰/۳۲۱۰۴۶	۶۵/۵۵۷۸۸	۲۵/۸۲۸۸۴
	SVR	۰/۸۶۵۹۱۲	۰/۳۶۶۱۸	۷۴/۷۷۴۱۷	۲۵/۳۸۰۲۳
Test	XGB	۰/۹۵۷۲۶۸	۰/۲۰۶۶۸۳	۴۳/۷۵۵۷۹	۱۸/۷۶۷۹۵
	Cat Boost	۰/۹۴۶۴۵۸	۰/۲۳۱۳۱۷	۴۸/۹۷۰۹۸	۱۹/۴۱۰۷۷
	Random Forest	۰/۹۳۱۴	۰/۲۶۱۸۹۴	۵۵/۴۴۴۳	۱۹/۳۱۷۳۹
	Extra Trees	۰/۹۲۶۸۲	۰/۲۷۰۴۶۹	۵۷/۲۵۹۵۷	۲۰/۲۰۹۰۳
	M5Prime	۰/۹۰۶۵۰۴	۰/۳۰۵۷۶۵	۶۴/۷۳۱۹۱	۲۳/۳۵۷۰۵
	K Neighbors	۰/۸۹۷۰۰۷	۰/۳۲۰۹۰۶	۶۷/۹۳۷۳۳	۲۴/۴۷۵۹۷
	Decision Tree	۰/۸۹۰۴۲۴	۰/۳۳۱۰۲۲	۷۰/۰۷۸۹۲	۲۴/۸۸۹۲۳
	MLP	۰/۸۸۵۸۵۸	۰/۳۳۷۴۴۸	۷۱/۴۳۹۴۸	۲۷/۶۲۷۱۷
	SVR	۰/۸۴۶۵۲۳	۰/۳۹۱۷۶۱	۸۲/۹۳۷۸۳	۲۷/۰۴۵۹۳

جدول (۴): مقایسه مدل‌های مورد بررسی در ایستگاه هیدرومتری پای پل با استفاده از معیارهای ارزیابی مختلف.

	Model	R^2	RRMSE	RMSE	MAE
Training	XGB	۰/۹۸۱۵۵۵	۰/۱۳۵۲۲۵	۳۱/۶۷۲۶۲	۱۷/۶۴۶۹۲
	Cat Boost	۰/۹۲۴۵۲۹	۰/۲۶۰۹۳۵	۶۱/۱۱۶۶	۲۵/۰۴۰۸۵
	Extra Trees	۰/۸۵۶۰۹۴	۰/۳۷۷۰۱۱	۸۸/۳۰۳۹۸	۲۸/۶۳۱۴
	Random Forest	۰/۸۴۴۰۸۲	۰/۳۸۷۹۲۸	۹۰/۸۶۱۰۱	۲۵/۵۶۲۸۹
	M5Prime	۰/۷۲۰۸۳	۰/۵۲۸۳۶۵	۱۲۳/۷۵۴۴	۳۸/۷۴۷۱۱
	K Neighbors	**۰/۷۱۰۷۴۶**	**۰/۵۳۷۸۲۳**	**۱۲۵/۹۶۹۷**	۴۰/۱۳۷۲۲
	Decision Tree	**۰/۷۰۹۷۲۶**	**۰/۵۳۸۷۷**	**۱۲۶/۱۹۱۶**	**۴۰/۳۹۵۸۸**
	MLP	۰/۷۰۹۳۳۶	۰/۵۳۹۱۲۸	۱۲۶/۲۷۵۴	۴۲/۴۰۰۲۱
	SVR	**۰/۶۸۸۷۴۷**	**۰/۵۵۷۹**	**۱۳۰/۶۷۲۲**	**۳۹/۲۳۶۷۲**
Test	XGB	۰/۷۸۱۵۸۷	۰/۴۶۷۳۴۲	۱۰۸/۸۷۰۴	۳۶/۴۵۸۴۵
	Cat Boost	۰/۷۵۹۳۱۲	۰/۴۹۰۵۸۷	۱۱۴/۲۸۵۴	۳۸/۱۶۲۳۳
	Extra Trees	۰/۷۴۲۶۰۵	۰/۵۰۷۳۲	۱۱۸/۱۸۳۵	۳۷/۴۹۸۹۲
	Random Forest	۰/۷۳۲۴۸۳	۰/۵۱۷۲۱۹	۱۲۰/۴۸۹۵	۳۵/۴۸۸۰۵
	Decision Tree	**۰/۷۱۳۸۴۶**	**۰/۵۳۴۹۳۴**	**۱۲۴/۶۱۶۳**	۴۰/۸۱۸۵۱
	K Neighbors	**۰/۷۱۳۱۷۹**	**۰/۵۳۵۵۵۷**	**۱۲۴/۷۶۱۴**	**۳۹/۷۷۴۳۴**
	M5Prime	۰/۷۱۰۶۶	۰/۵۳۷۹۰۳	۱۲۵/۳۰۸۱	۳۹/۴۳۴۴۴
	MLP	۰/۷۰۲۶۶۴	۰/۵۴۵۲۶۳	۱۲۷/۰۲۲۶	۴۲/۴۲۷۱۳
	SVR	**۰/۷۰۰۸۳۶**	**۰/۵۴۶۹۵۹**	**۱۲۷/۴۱۷۷**	**۳۸/۵۲۷۷**

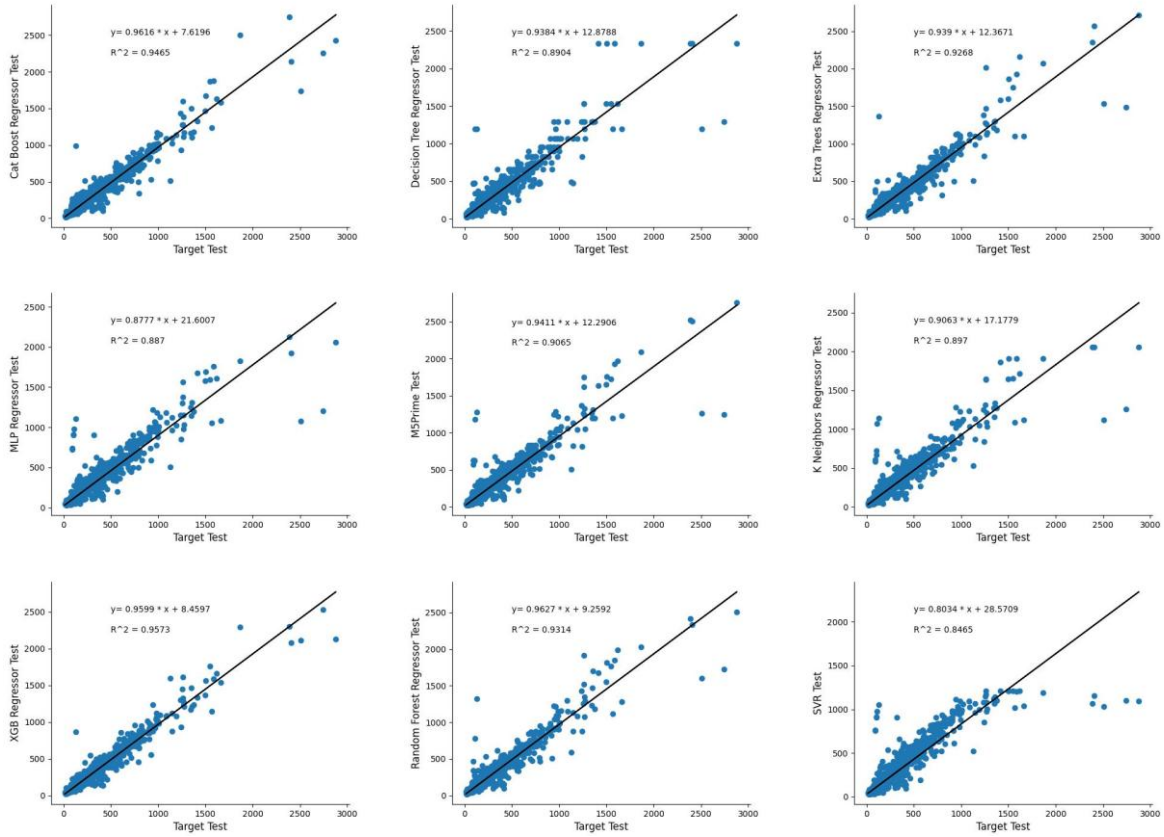


شکل (۳): تخمین جریان روزانه با استفاده از مدل‌های مختلف یادگیری ماشین در ایستگاه عبدالخان.

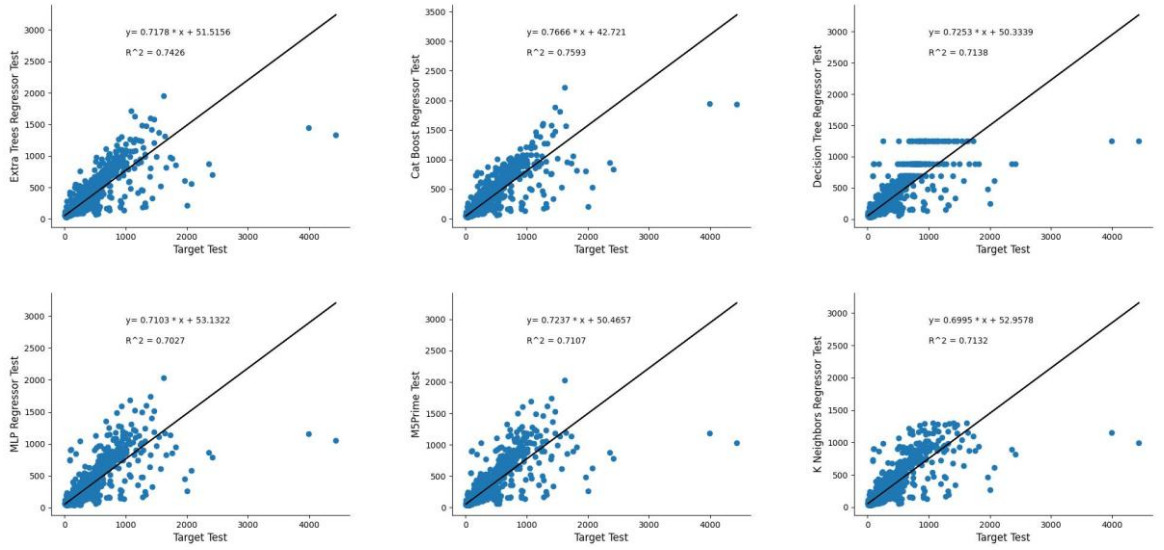


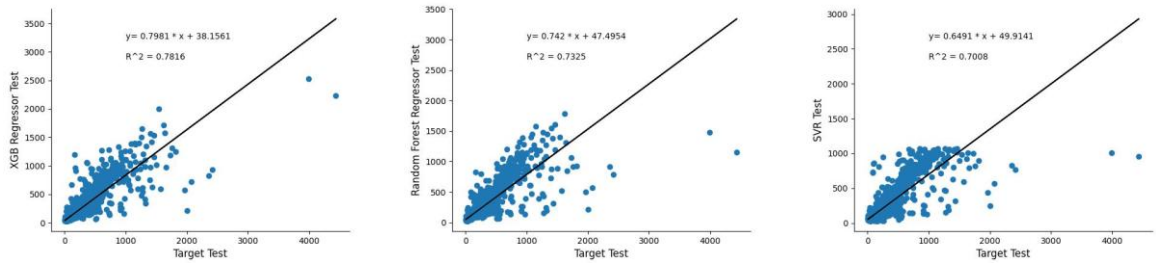
شکل (۴): تخمین جریان روزانه با استفاده از مدل‌های مختلف یادگیری ماشین در ایستگاه پای پل.

شکل‌های ۵ و ۶ نمودارهای پراکنش مقادیر جریان روزانه رودخانه کرخه در دو ایستگاه هیدرومتری عبدالخان و پای پل در دوره آزمون را نشان می‌دهند. مطابق شکل‌های مذکور، مشاهده می‌شود که داده‌های تخمین زده شده با استفاده از دو مدل Xgboost و Catboost در هر دو ایستگاه مذکور دارای انطباق خوبی با داده‌های مشاهداتی نسبت به سایر مدل‌ها در مرحله صحت‌سنجی دارند و مقادیر مفقود جریان روزانه با همبستگی بالایی تخمین زده شده است. البته مدل Xgboost نسبت به مدل Catboost دقیق‌تر بوده و مقادیر جریان روزانه رودخانه را در ایستگاه عبدالخان، با داشتن ضریب تعیین ۰/۹۵ نسبت به ۰/۹۴ مدل Catboost با دقت بیشتری تخمین زده است. با توجه به بالا بودن مقدار R^2 مدل‌های مذکور روی داده‌های آموزشی و آزمایشی، مدل در حالت تعادل می‌باشد. بنابراین، حالت بیش‌برازش و کم‌برازش اتفاق نیفتاده است. در ایستگاه پای پل نیز نتایج نشان‌دهنده تطابق بیشتر داده‌های تخمین زده شده با داده‌های مشاهداتی توسط مدل Xgboost نسبت به مدل Catboost می‌باشد.



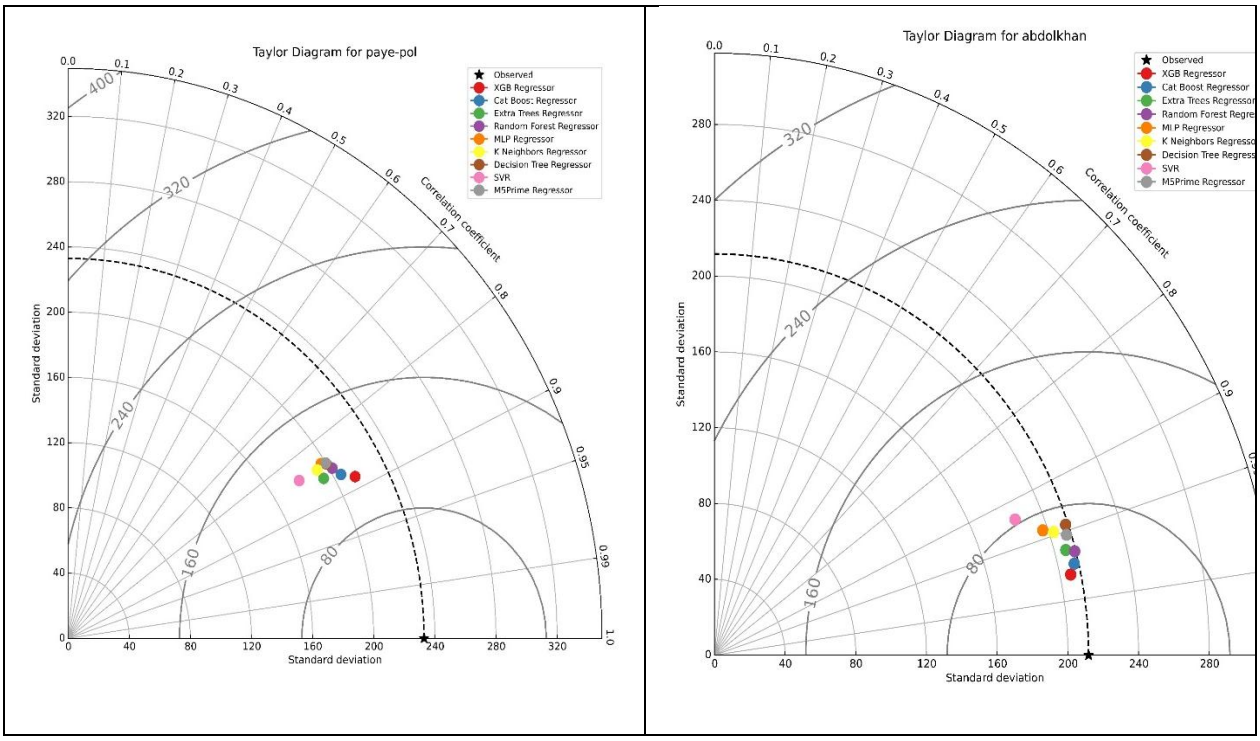
شکل (۵): نمودارهای پراکنش مقادیر جریان روزانه رودخانه کرخه در ایستگاه هیدرومتری عبدالخان در دوره آزمون.





شکل (۶): نمودارهای پراکنش مقادیر جریان روزانه رودخانه کرخه در ایستگاه هیدرومتری پای پل در دوره آزمون.

برای بررسی بیشتر و مقایسه خروجی‌های مدل‌سازی در یک نمودار، از دیاگرام تیلور استفاده شد. در شکل ۷، عملکرد مدل‌های یادگیری ماشین Xgbooxt، Catboost، Extra trees، Random forest، M5، MLP، K-NN، Decision tree و SVR نشان داده شده است. بیشترین همبستگی و کمترین خطا مربوط به مدل Xgboost است. لذا بر اساس بررسی‌ها از منظر مختلف، در هر دو ایستگاه هیدرومتری عبدالخان و پای پل، این گونه استنباط می‌گردد که بهترین عملکرد مربوط به مدل Xgboost و پس از آن Catboost می‌باشد و مدل Xgboost نزدیکترین خروجی‌ها نسبت به داده‌های مشاهداتی را شبیه‌سازی می‌کند که این نشان‌دهنده عملکرد بهتر و دقت بیشتر در تخمین داده‌های مفقود جریان رودخانه است.



شکل (۷): نمودار تیلور برای دو ایستگاه هیدرومتری عبدالخان و پای پل.

نتیجه‌گیری

داده‌های جریان روزانه رودخانه‌ها برای مدیریت منابع آب، برنامه‌ریزی هیدرولوژیک و پیش‌بینی سیلاب‌ها بسیار حیاتی هستند. متأسفانه، این آمار که می‌تواند مبنای ارزشمندی در تحقیقات باشد گاه دارای داده‌های مفقود است. داده‌های مفقود می‌توانند تحلیل‌ها را به چالش بکشند و دقت مدل‌های پیش‌بینی را کاهش دهند. از این رو، تخمین دقیق داده‌های مفقود امری ضروری است. در این پژوهش، به منظور تخمین داده‌های مفقود جریان روزانه رودخانه کرخه، با استفاده از طیف وسیعی از روش‌های یادگیری جمعی و ماشینی نظارت شده و همچنین تنظیم

فراپارامترهای الگوریتم‌ها با استفاده از روش Optuna، بهترین مدل جهت تخمین داده‌های مفقود جریان روزانه رودخانه ارائه گردید. روش بهینه سازی Optuna باعث شد دقت و کارایی مدل‌ها را بهبود ببخشد و به توسعه راهکارهای بهینه‌تر برای حل مسائل پیچیده هیدرولوژیک کمک کند. نتایج با استفاده از معیارهای ارزیابی MAE، RMSE، RRMSE، R^2 و همچنین استفاده از نمودار تیلور به عنوان ابزاری برای مقایسه عملکرد مدل‌ها بسیار مؤثر بوده است. نتایج نشان داد که مدل Xgboost با کمترین مقدار MAE برابر ۱۸/۷۶ و بیشترین مقدار R^2 برابر ۰/۹۵ بهترین عملکرد را در بین مدل‌ها برای تخمین جریان روزانه رودخانه کرخه در ایستگاه هیدرومتری عبدالخان داشته است. در ایستگاه پای پل نیز پس از مدل Xgboost، مدل Catboost عملکرد خوبی جهت تخمین داده‌های مفقود داشت. بقیه مدل‌ها نیز عملکرد متوسطی داشته‌اند. به‌کارگیری تکنیک‌های داده‌کاوی جهت برآورد تخمین جریان روزانه رودخانه، مخصوصاً در کشورهای در حال توسعه که دسترسی به اطلاعات و داده‌های هیدرولوژیک با مشکل و محدودیت مواجه هستند، جهت اقدامات مدیریتی مناسب که سبب کاهش خسارت‌ها و تلفات ناشی از بحران‌های حدی می‌گردد، مفید می‌باشد. با توجه به کارایی بالای روش Optuna در بهینه‌سازی فراپارامترهای مدل‌های یادگیری ماشین در این پژوهش، پیشنهاد می‌شود جهت بهینه‌سازی فراپارامترهای مدل‌ها در مقیاس‌های زمانی دیگر مانند داده‌های ماهانه یا سالانه نیز به کار گرفته شوند. این امر می‌تواند برای مدل‌سازی جریان‌های بلندمدت و تحلیل‌های هیدرولوژیک مفید باشد. مهم‌ترین نتایج به‌دست آمده در پژوهش حاضر به شرح زیر است:

۱. مدل‌های یادگیری جمعی مانند Xgboost و Catboost توانسته‌اند با ترکیب نقاط قوت روش‌های مختلف، دقت بالاتری در تخمین داده‌های مفقود داشته باشند. این مدل‌ها برخلاف روش‌های تکی مانند SVR و درخت تصمیم، توانسته‌اند خطای پیش‌بینی را کاهش دهند و تخمین‌هایی دقیق‌تر و قابل اعتمادتر ارائه دهند.

۲. در ایستگاه عبدالخان، با وجود داشتن داده‌های مفقود بیشتر نسبت به ایستگاه پای پل، دقت مدل‌های استفاده شده بیشتر بود، که علت آن نزدیکی ایستگاه عبدالخان به ایستگاه هیدرومتری حمیدیه است. این نشان‌دهنده اهمیت ویژگی‌های مکانی در کیفیت خروجی مدل‌های یادگیری ماشین است.

۳. در ایستگاه هیدرومتری پای پل، مدل Xgboost با وجود فاصله زیاد از ایستگاه هیدرومتری حمیدیه، به عنوان ایستگاه همسایه، توانایی خوبی در برآورد داده‌های مفقود جریان روزانه رودخانه، در بین مدل‌های Catboost، Extra trees، Random forest، M5، MLP، K-NN، Decision Tree، Cart و SVR داشته است.

٤. مدل‌های مبتنی بر Boosting می‌توانند بر چالش‌های فاصله مکانی و داده‌های محدود غلبه کنند.

۵. مدل Catboost در ایستگاه عبدالخان و پای پل به ترتیب با ۱۱ درصد و ۵ درصد دقت کمتر نسبت به مدل Xgboost توانست دومین جایگاه را در بین مدل‌های بررسی شده کسب کند، که این امر نشان‌دهنده قدرت مدل CatBoost در تخمین داده‌های هیدرولوژیک است.

٦. مدل SVR بر اساس معیارهای ارزیابی MAE، RMSE، RRMS، R^2 و نمودار تیلور، ضعیف‌ترین عملکرد را در بین ۹ مدل بررسی شده داشته است.

۷. مدل‌های یادگیری جمعی می‌توانند ضعف روش‌های انفرادی مانند رگرسیون بردار پشتیبان و ... را پوشش دهند تا نتایج بهتری در تخمین جریان رودخانه حاصل شود.

۸. استفاده از روش بهینه‌سازی Optuna برای تنظیم فراپارامترهای مدل‌های یادگیری ماشین، نقش مهمی در بهبود عملکرد مدل‌ها داشته است.

منابع

آهنی، ع. و شوریان، م. ۱۳۹۶. پیش‌بینی جریان ماهیانه رودخانه با استفاده از مدل‌های داده مبنا. تحقیقات منابع آب ایران. سال سیزده، شماره دو، ص ۲۰۷-۲۱۴.

بوستانی، م.، کرمی، ح.، موسوی، س. ف.، و فرزین، س. ۱۳۹۸. بررسی ارتباط بین شاخص‌های نظریه آشوب در رفتارنگاری جریان رودخانه‌ای در مقیاس‌های زمانی کوتاه‌مدت. نشریه علمی پژوهشی مهندسی آبیاری و آب ایران، سال نهم، شماره چهار، ص ۹۸-۱۱۶.

دارابی، ف.، نجفی نژاد، ع.، پورقاسمی، ح. ر. و سعدالدین، ا. ۱۴۰۳. پیش‌بینی اثر اقدامات بیولوژیک بر سیل‌خیزی حوزه آبخیز بهشت‌آباد با استفاده از روش‌های یادگیری ماشین. مدیریت جامع حوزه‌های آبخیز. 10.22034/iwm.2024.2032264.1159

سیدیان، س. م.، سلیمانی، م. و کاشانی، م. ۱۳۹۳. پیش‌بینی دبی جریان رودخانه با استفاده از داده‌کاوی و سری زمانی. اکوهیدرولوژی. سال یک، شماره سه، ص ۱۶۷-۱۷۹.

غفاری، غ. ع. و وفاخواه، م. ۱۳۹۲. شبیه‌سازی فرآیند بارش- رواناب با استفاده از شبکه عصبی مصنوعی و سیستم فازی-عصبی تطبیقی (مطالعه موردی: حوزه آبخیز حاجی‌قوشان). پژوهشنامه مدیریت حوزه آبخیز. سال چهار، شماره هشت، ص ۱۳۶-۱۲۰.

محمدی، س.، حسن پور، ف.، شریف آذری، س. و فروغی، ف. ۱۴۰۰. ارزیابی روش‌های رگرسیونی نوین جهت تخمین بار رسوبی معلق در رودخانه سیستان. نشریه علمی پژوهشی مهندسی آبیاری و آب ایران. سال دوازده، شماره دو، ص ۱-۱۵.

میرنورالهی، ع.، کرمی، ح.، فرزین، س. و عامری، م. ۱۴۰۱. بررسی عملکرد ماشین‌های یادگیری در تخمین ضریب دبی آبگذری آبگیرهای کفی با روزنه دایره‌ای. نشریه علمی پژوهشی مهندسی آبیاری و آب ایران، سال دوازده، شماره چهار، ص ۲۱-۴۱.

یوسفی، ح.، یونسی، ح، ا.، داودی مقدم، د.، ارشیا، آ. و شمسی، ز. ۱۴۰۱. تعیین پتانسیل سیل با استفاده از مدل‌های یادگیری ماشین CART، GLM و GAM مطالعه موردی: حوضه کشکان. نشریه علمی پژوهشی مهندسی آبیاری و آب ایران، سال دوازده، شماره چهار، ص ۱۰۵-۸۴.

Ahadiyan, J. (2016). Application of ANFIS adaptive system to estimate the potential consolidation of clay soils. Journal of Modeling in Engineering, 14(45), 17-31.

Ahadiyan, J., Kiani, S., Asiaban, P., Azizi Nadian, H., & Omidvarinia, M. (2023). Optimizing the dimensions of the agricultural water transfer system from the Karun 3 dam to the northeastern cities of Khuzestan province. Journal of New Approaches in Water Engineering and Environment, 1(2), 112-126.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. 2019. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

Bagherian Marzouni, M., Akhoundali, A. M., Moazed, H., Jaafarzadeh, N., Ahadian, J., & Hasoonizadeh, H. (2014). Evaluation of Karun river water quality scenarios using simulation model results. International Journal of Advanced Biological and Biomedical Research, 2(2), 339-318.

Boustani, M., Farzin, S., & Mousavi, S. F. (2025). Estimation of R-Vine copula parameters in multivariate flood frequency analysis using arithmetic optimization algorithm and comparing the performance with genetic algorithm. Water Resources Management, 1-19.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on computational learning theory (pp. 144-152).

Breiman, L. 2001. Random forests. Machine Learning, 45, 5-32.

Breiman, L., Friedman, J., & Olshen, R. A. 2017. Classification and regression trees. Routledge.

Chemura, A., Rwasoka, D., Mutanga, O., Dube, T., & Mushore, T. 2020. The impact of land-use/land cover changes on water balance of the heterogeneous Buzi sub-catchment, Zimbabwe. Remote Sensing Applications: Society and Environment, 18, 100292.

Chen, T., & Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

Daryaei, M., Kashefipour, S. M., Ahadian, J., & Ghobadian, R. (2010). Modeling the compression index of fine soils using artificial neural network and comparison with the other empirical equations. Journal of Water and Soil, 24(4), 659-667.

Dariane, A. B., & Borhan, M. I. 2024. Comparison of classical and machine learning methods in estimation of missing streamflow data. Water Resources Management, 38(4), 1453-1478.

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., ... & Xiang, Y. 2018. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy Conversion and Management, 164, 102-111.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189-1232.

Geurts, P., Ernst, D., & Wehenkel, L. 2006. Extremely randomized trees. Machine Learning, 63, 3-42.

Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. 2021. Catboost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166, 120658.

Katipoğlu, O. M., & Sarıgöl, M. 2023. Prediction of flood routing results in the Central Anatolian region of Türkiye with various machine learning models. Stochastic Environmental Research and Risk Assessment, 37(6), 2205-2224.

Khampuengson, T., & Wang, W. 2023. Novel methods for imputing missing values in water level monitoring data. Water Resources Management, 37(2), 851-878.

Khoramipoor, Z., Valikhan Anaraki, M., & Farzin, S. 2024. A new approach in flood routing based on the integration of bayes theory, support vector machine and meta-heuristic optimization algorithm. Iranian Journal of Irrigation & Drainage, 18(3), 409-420.

Krysanova, V., & White, M. 2015. Advances in water resources assessment with SWAT-an overview. Hydrological Sciences Journal, 60(5), 771-783.

Latifoğlu, L., & Canpolat, Ü. (2022). Prediction of daily streamflow data using ensemble learning models. The European Journal of Research and Development, 2(4), 356-371.

Li, Y., Liang, Z., Hu, Y., Li, B., Xu, B., & Wang, D. 2020. A multi-model integration method for monthly streamflow prediction: Modified stacking ensemble strategy. Journal of Hydroinformatics, 22(2), 310-326.

Minns, A. W., & Hall, M. J. 1996. Artificial neural networks as rainfall-runoff models. Hydrological Sciences Journal, 41(3), 399-417.

Mohammadi, M., Vagharfard, H., Mahdavi Najafabadi, R., Daneshkar Arasteh, P., & Nazemosadat, M. J. 2021. Rainfall-runoff modelling of coastal watersheds near Hormuz Strait using data mining. Iranian Journal of Soil and Water Research, 52(2), 313-327.

Molnar, C. 2020. Interpretable machine learning. Lulu. com.

Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., & Liu, J. 2020. Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. Journal of Hydrology, 586, 124901.

Quinlan, J. R. 1992. Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348).

Razavi, T., & Coulibaly, P. 2013. Streamflow prediction in ungauged basins: Review of regionalization methods. Journal of Hydrologic Engineering, 18(8), 958-975.

Samadi, M., Bahremand, A., & Fathabadi, A. 2019. The Boustan Dam monthly inflow forecasting using data-driven and ensemble models in the Golestan Province. Watershed Engineering and Management, 11(4), 1044-1058.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406, 109-120.

Sharma, V., & Yuden, K. 2021. Imputing missing data in hydrology using machine learning models. International Journal of Engineering Research and Technology, 10, 78-82.

Sharififard, E., Azizipour, M., Ahadiyan, J., & Haghighi, A. (2024). Determination of creep function coefficients of viscoelastic pipes using a transient-guided machine learning model. AQUA—Water Infrastructure, Ecosystems and Society, 73(11), 2132-2149.

Sumayli, A. 2023. Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models. Arabian Journal of Chemistry, 16(7), 104833.

Taylor, K. E. 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106(D7), 7183-7192.

Terzi, Ö., Küçüksille, E. U., Baykal, T., & Taylan, E. D. 2023. Deep and machine learning for daily streamflow estimation: A focus on LSTM, RFR and Xgboost. Water Practice & Technology, 18(10), 2401-2414.

Varga, M., Balogh, S., & Csukas, B. 2016. GIS based generation of dynamic hydrological and land patch simulation models for rural watershed areas. Information Processing in Agriculture, 3(1), 1-16.

Venkatesan, E., & Mahindrakar, A. B. 2019. Forecasting floods using extreme gradient boosting–a new approach. International Journal of Civil Engineering and Technology, 10(2), 1336-1346.

Xie, T., Chen, L., Yi, B., Li, S., Leng, Z., Gan, X., & Mei, Z. 2024. Application of the improved k-nearest neighbor-based multi-model ensemble method for runoff prediction. Water, 16(1), 69.

Yohaness, Y. 1999. Classification and regression tree: An introduction. Research Institute of Washington, DC.

Zhang, Y., Zhao, Z., & Zheng, J. 2020. Catboost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. Journal of Hydrology, 588, 125087.