Iranian Journal of Soil and Water Research ISNN: 2423-7833
Homepage: http://ijswr.ut.ac.ir

The combination of dimensionality reduction methods and machine learning algorithms in the optimization of Maroon River water quality prediction

Fereshteh Sayahi[1] | Laleh Divband Hafshejani[2]✉ | Parvaneh Tishehzan[3] | Hamid Abdolabadi[4]

1. Department of Environmental Engineering, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran. E-mail: f.sayahi75@gmail.com
2. Corresponding Author, Department of Environmental Engineering, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran. E-mail: l.divband @scu.ac.ir
3. Department of Environmental Engineering, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran: partishezan@scu.ac.ir
4. Department of Environmental Engineering, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran: h.abdolabadi@scu.ac.ir

Article Info

Article type: Research Article

Article history:

Received: May. 7, 2024

Revised: July. 12, 2024

Accepted: Aug. 5, 2024

Published online: Nov. 2024

Keywords:
Nitrate,
Linear Regression,
Random Forest,
Evaluation Criteria.

ABSTRACT

Water resources face challenges such as climate change and human activities. Sustainable water management is extremely important to solve this problem. More and more people are using artificial intelligence, especially machine learning, to predict and manage water quality. These AI methods are excellent at identifying patterns in water data and improving water quality management. This study examines the water quality of the Maroon River using a combination of factor analysis and machine learning. Data on various water quality parameters were collected from three stations over a period of ten years and the water quality index was calculated. Then, different machine learning algorithms were used to predict the water quality index. In a further step, factor analysis was performed to extract the important features of the input for the optimal algorithm. The performance of the studied algorithms was determined at each step using evaluation criteria. The results showed that in the first step, the Random Forest algorithm (R^2 (0.78), RMSE (2.65)) had the best performance in predicting water quality index. It was also found that among the three algorithms studied, nitrate is the most important input parameter, while acidity is the least important. By reducing the number of inputs to 3 important parameters, the performance of the Random Forest algorithm (R^2 (0.74), RMSE (2.86)) almost reached the level of 8 input parameters. Combining insights from factor analysis and feature importance analysis can provide a more comprehensive understanding of the complex relationships among water quality parameters and help develop more effective water management.

Cite this article Sayahi, F., Divband Hafshejani, L., Tishehzan, P., & Abdolabadi, H., (2024) The combination of dimensionality reduction methods and machine learning algorithms in the optimization of Maroon River water quality prediction, *Iranian Journal of Soil and Water Research*, 55 (9), 1601-1615. https://doi.org/10.22059/ijswr.2024.376275.669708

 Publisher: The University of Tehran Press.

DOI: https://doi.org/10.22059/ijswr.2024.376275.669708

EXTENDED ABSTRACT

Introduction:

In today's world, water resources have attracted much attention due to their unique importance. These resources are of great value as one of the vital bases for human life, environmental protection and economic development. With population increase, climate change and human pressures, water resources are facing many challenges and threats, especially in dry areas. These challenges include reducing water quality and quantity, destroying water resources, and creating serious problems for freshwater consumption. Therefore, the importance of investigating and sustainable management of water resources is of particular importance. In this regard, the use of artificial intelligence methods, especially machine learning, is increasingly used in predicting and modelling water quality and water resources management. Due to their ability to detect patterns and complex relationships in water quality data, these methods are considered effective tool for improving water quality management and maintenance.

Materials and Methods:

The present study examines the water quality of the Maroon River, one of the most important rivers in Iran, which plays an important role in the development of urban and rural areas. The data used include parameters such as temperature, biochemical oxygen demand, phosphate... for 10 years have been collected from different stations. In the first step, these data have been used as inputs for forecasting models. Then, dimension reduction methods such as factor analysis have been used to extract important features. In the next step, different machine learning algorithms such as Linear Regression،Random Forest ،Extra Trees Light Gradient Boosting Machine have been used to predict the water quality index, and the performance of the algorithms was evaluated using criteria such as root mean square error and coefficient of determination.

Results and Discussion:

The p-value of Bartlett's test in this research was close to zero and it can be concluded that there is a significant correlation between the variables and the data are suitable for factor analysis and dimension reduction. The values of the variance inflation coefficient for the water quality parameters used in this research showed that total coliform and phosphate variables have little colinearity with other independent variables. The prediction results of the water quality index using the 8 studied parameters as input showed that the random forest and regression algorithms showed the highest and lowest agreement with the real data, respectively. Because the regression algorithm uses a straight line to predict the dependent variable's values based on the independent variables and performs poorly in complex problems with non-linear interactions. The results also showed that nitrate is the most important input parameter and acidity is less important for the three studied algorithms.

Conclusion:

By combining the insights obtained from factor analysis and feature importance analysis, researchers can better understand the complex relationships between water quality parameters and create more effective strategies for water management and pollution control.

Author Contributions

Fereshteh Sayahi: Design, Analysis, and Interpretation of data Writing- Original draft preparation, Visualization. Laleh Divband Hafshejani: Conceptualization, Methodology, Design, Revision of the manuscript and Editing. Parvaneh Tishehzan: Design, Revision of the manuscript and Editing. Hamid Abdolabadi: Analysis and Interpretation of data.

Data Availability Statement

Data can be sent from the corresponding author by email upon request.

Acknowledgements

We are grateful to the Research Council of Shahid Chamran University of Ahvaz for financial support (GN SCU.WE1402.47794).

Ethical considerations

The authors avoided data fabrication, falsification, plagiarism, and misconduct.

مجله تحقیقات آب و خاک ایران، دوره ۵۵، شماره ۹ شاپا: ۷۸۳۳-۲۴۲۳
Homepage: http://ijswr.ut.ac.ir

ترکیب روش‌های کاهش ابعاد و الگوریتم‌های یادگیری ماشین در بهینه‌سازی پیش‌بینی کیفیت آب رودخانه مارون

فرشته سیاحی[۱] | لاله دیوبند هفشجانی[۲]✉ | پروانه تیشه‌زن[۳] | حمید عبدل آبادی[۴]|

۱. گروه مهندسی محیطزیست، دانشکده مهندسی آب و محیطزیست، دانشگاه شهید چمران اهواز، اهواز، ایران. رایانامه: f.sayahi75@gmail.com
۲. نویسنده مسئول، گروه مهندسی محیطزیست، دانشکده مهندسی آب و محیطزیست، دانشگاه شهید چمران اهواز، اهواز، ایران. رایانامه: l.divband@scu.ac.ir
۳. گروه مهندسی محیط زیست، دانشکده مهندسی آب و محیطزیست، دانشگاه شهید چمران اهواز، اهواز، ایران. رایانامه: partishezan@scu.ac.ir
۴. گروه مهندسی محیط زیست، دانشکده مهندسی آب و محیطزیست، دانشگاه شهید چمران اهواز، اهواز، ایران. رایانامه: h.abdolabadi@scu.ac.ir

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۳/۸/۲۲
تاریخ بازنگری: ۱۴۰۳/۴/۲۲
تاریخ پذیرش: ۱۴۰۳/۵/۱۵
تاریخ انتشار: آذر ۱۴۰۳

واژه‌های کلیدی:
نیترات،
رگرسیون خطی،
جنگل تصادفی،
معیارهای ارزیابی.

چکیده

منابع آب برای زندگی انسان، رشد اقتصادی و حفظ محیطزیست حیاتی میباشند، اما با چالشهایی مانند تغییرات آب و هوایی و فعالیت‌های انسانی، به‌ویژه در مناطق خشک مواجه هستند. برای رفع این مشکل، مدیریت پایدار آب بسیار مهم است. هوش مصنوعی، به‌ویژه یادگیری ماشین، به طور فزایندهای برای پیشبینی و مدیریت کیفیت آب استفاده میشود. این روش‌های هوش مصنوعی در تشخیص الگوها در داده‌های آب عالی هستند و به بهبود مدیریت کیفیت آب کمک می‌کنند. بنابراین در این مطالعه به بررسی کیفیت آب رودخانه مارون با استفاده از ترکیب روش تحلیل عاملی و یادگیری ماشین پرداخته شد. داده‌های ۱۰ ساله پارامترهای مختلف کیفیت آب در سه ایستگاه جمع‌آوری گردید و شاخص کیفیت آب ایران برای هر سری داده محاسبه شد. سپس الگوریتم‌های مختلف یادگیری ماشین برای پیش‌بینی شاخص کیفیت آب به کار گرفته شده‌اند. در مرحله بعد تحلیل عاملی برای استخراج ویژگی‌های مهم ورودی به الگوریتم بهینه استفاده گردید. قابل‌ذکر است در هر مرحله عملکرد الگوریتمهای موردمطالعه با استفاده از معیارهای ارزیابی تعیین شد. نتایج نشان داد که در مرحله اول الگوریتم جنگل تصادفی (R^2 (0.78) و (RMSE (2.65) بهترین عملکرد را در پیش‌بینی شاخص کیفیت آب داشت. همچنین مشخص شد که نیترات مهمترین پارامتر ورودی و اسیدیته کم‌اهمیت‌ترین پارامتر برای سه الگوریتم موردمطالعه است. قابل‌ذکر است در حالتی که تعداد ورودیها به ۳ پارامتر با اهمیت کاهش یافت، عملکرد الگوریتم جنگل تصادفی (R^2 (0.74) و (RMSE (2.86) تقریبا مشابه با ۸ پارامتر ورودی بود. ترکیب بینش‌های حاصل از تحلیل عاملی و تحلیل اهمیت ویژگی‌ها می‌تواند درک جامع‌تری از روابط پیچیده بین پارامترهای کیفیت آب ارائه دهد و به ایجاد استراتژی‌های موثرتر برای مدیریت آب و کنترل آلودگی کمک کند.

استناد: سیاحی؛ فرشته، دیوبند هفشجانی؛ لاله، تیشه زن؛ پروانه، عبدل آبادی؛ حمید، (۱۴۰۳) ترکیب روش‌های کاهش ابعاد و الگوریتم‌های یادگیری ماشین در بهینه سازی پیش‌بینی کیفیت آب رودخانه مارون، *مجله تحقیقات آب و خاک ایران*، ۵۵ (۹)، ۱۶۱۵-۱۶۰۱.
https://doi.org/10.22059/ijswr.2024.376275.669708

ناشر: مؤسسه انتشارات دانشگاه تهران.

DOI: https://doi.org/10.22059/ijswr.2024.376275.669708

مقدمه

در دنیای امروز، منابع آبی به دلیل اهمیت بی‌نظیری که دارند، توجه بسیاری را به خود جلب کرده‌اند. با توجه به اینکه آب به عنوان یکی از منابع حیاتی برای زندگی انسان، حفظ محیط‌زیست و توسعه اقتصادی مورد ارزش قرار می‌گیرد (Ahmed et al., 2019; Watkins, 2006) ، با افزایش جمعیت، تغییرات اقلیمی و فشارهای انسانی، منابع آبی با چالش‌ها و تهدیدات فراوانی مواجه هستند. علاوه بر این، در مناطق خشک، کاهش کیفیت و کمیت آب و نابودی منابع آبی نتیجه‌ای از این تهدیدات است. بنابراین، درک اهمیت منابع آب و اتخاذ اقدامات مؤثر برای مدیریت پایدار آنها ضروری است تا بتوان به توسعه پایدار و حفظ محیط‌زیست دست یافت (Watkins, 2006). با وجود این موضوعات، متأسفانه، به دلیل تغییرات در اکوسیستم‌ها، صنعتی شدن، افزایش کشاورزی و شهرنشینی، کیفیت این منابع به شدت کاهش یافته و برای استفاده در بخش‌های مختلف قابل استفاده نیستند. همچنین، بدون دسترسی کافی به آب شیرین، مشکلات جدی به وجود می‌آید (Jatnika et al., 2021). یکی از منابع آبی رودخانه‌ها هستند که نقش اساسی در محیط‌زیست انسانی و توسعه شهری دارند و به عنوان یکی از منابع آب حیاتی برای آبیاری، نیازهای صنعتی و سایر مصارف محسوب می‌شوند (Chen et al., 2021; Chen et al., 2023). رودخانه‌ها، به دلیل ویژگی‌هایی که دارند، همواره در معرض تغییر و تحول قرار می‌گیرند و در واقع ماهیت دینامیکی‌شان، آن‌ها را در برابر تأثیرات نامطلوب آلودگی محیطی آسیب‌پذیرتر می‌کند (Ahmed et al., 2019). آلودگی رودخانه زمانی رخ می‌دهد که آلاینده‌ها وارد رودخانه‌ها شوند و بر مصارف آن‌ها تأثیر منفی بگذارند، که این موضوع یک مسئله مهم زیست‌محیطی است و می‌تواند اثرات شدیدی بر اکوسیستم‌های آبی، سلامت انسان و اقتصاد داشته باشد. بنابراین، ترکیبی از اقدامات نظارتی، راه‌حل‌های تکنولوژیکی و آموزش عمومی برای مدیریت و حفظ پایداری رودخانه‌ها و منابع آب حائز اهمیت است (Li et al., 2023).

اخیراً روش‌های هوش مصنوعی (AI) به طور فزاینده‌ای برای پیش‌بینی، مدل‌سازی و بهینه‌سازی کیفیت آب مورد استفاده قرار گرفته‌اند. هوش‌مصنوعی منجر به ایجاد یک ساختار ریاضی انعطاف‌پذیر می‌شود که قادر به شناسایی روابط غیرخطی و پیچیده بین داده‌های ورودی و خروجی است (Ahmed et al., 2019). یادگیری ماشین به‌عنوان یک زیرشاخه از هوش مصنوعی، به‌عنوان یک ابزار مؤثر برای استخراج مدل‌های پیش‌بینی از داده‌ها شناخته می‌شود. این ابزار، امکان یادگیری ویژگی‌های با ابعاد بالا و روابط غیرخطی را فراهم می‌کند (Haggerty et al., 2023). از روش‌های مختلفی مانند طبقه‌بندی، رگرسیون و خوشه‌بندی بهره می‌برد و به رایانه‌ها امکان می‌دهد با دقت و سرعت بالا اشیاء را طبقه‌بندی کنند (Zhu et al., 2022; Ali et al., 2023). قابل‌ذکر است که استفاده از یادگیری ماشین نسبت به انسان، به دلیل کارایی بالاتر و نتایج مستقل‌تر، مزیت دارد (Divband Hafshejani et al., 2022). این روش‌ها در حوزه‌های مختلف مدیریت آب و مدیریت زیست‌محیطی نیز مؤثر هستند. الگوریتم‌های یادگیری ماشینی، به دلیل توانایی مدیریت روابط پیچیده بین متغیرها و سازگاری با شرایط متغیر، قادر به ارائه پیش‌بینی‌های دقیقی از پارامترهای کیفیت آب هستند (Ahmed et al., 2019). استفاده از الگوریتم‌های یادگیری ماشینی در مسائل کیفیت آب می‌تواند به طور بالقوه روابط و الگوهای پیچیده‌تری را در داده‌ها که ممکن است از طریق روش‌های آماری سنتی آشکار نباشد، آشکار کند (Krishnan & Manikandan., 2024). این الگوریتم‌ها می‌توانند حجم زیادی از داده‌ها را به سرعت پردازش کرده و پیش‌بینی کیفیت آب را به صورت زمان واقعی یا حداقل تقریبی فراهم کنند (Azrour et al., 2022). همچنین، با خودکارسازی فرآیند پیش‌بینی، نیاز به تجزیه و تحلیل دستی داده‌ها را کاهش داده و صرفه‌جویی در زمان و منابع را فراهم می‌کنند (Zhu et al., 2022). از این رو، استفاده از الگوریتم‌های یادگیری ماشینی برای پیش‌بینی کیفیت آب، به منظور مدیریت مؤثر منابع آب بسیار حیاتی و مهم است. تحقیقات انجام‌شده در این راستا نشان‌دهنده اهمیت موضوع نیز است. در تحقیقی انجام‌شده توسط Schäfer et al., 2022، کیفیت آب در یک رودخانه واقع در جنوب شرقی انگلستان با استفاده از روش یادگیری ماشین مورد بررسی قرار گرفت. آن‌ها از داده‌های هدایت الکتریکی و دما استفاده کرده و از دو الگوریتم درخت تصمیم و رگرسیون برای تجزیه و تحلیل داده‌ها استفاده کردند. نتایج نشان داد که الگوریتم درخت تصمیم عملکرد بهتری داشته و قادر به تخمین کیفیت آب با دقت بالا و خطای کمتر از ۱ درصد است. در مطالعه انجام‌شده توسط Deng et al., 2021، پیش‌بینی کیفیت آب در بندر طلوع واقع در هنگ‌کنگ با استفاده از یادگیری ماشین بررسی شد. این تحقیق از داده‌های بیش از ۳۰ ساله آماری استفاده کرده است. برای شبیه‌سازی و مقایسه داده‌ها، از روش‌های یادگیری ماشین، از جمله شبکه‌های عصبی مصنوعی و ماشین بردار پشتیبان استفاده شده است. نتایج این تحقیق نشان داد که شبکه‌های عصبی مصنوعی پاسخ سریع‌تری ارائه می‌دهند، در

حالی که ماشین بردار پشتیبان دقیق‌تر عمل می‌کند، اما زمان‌برتر است.

شناسایی مهم‌ترین پارامترهای موثر بر کیفیت آب و کاهش تعداد پارامترهای موردنیاز برای تعیین وضعیت کیفیت آب به مدیران منابع آب اجازه می‌دهد تا ماهیت پیچیده مسائل کیفیت آب را درک کرده و اقداماتی را برای حفظ کیفیت آب اولویت‌بندی کنند (Sharma et al., 2021) تجزیه و تحلیل عاملی به طور گسترده‌ای به عنوان یک تکنیک آماری برای ارزیابی کیفیت آب استفاده می‌شود. در این روش مهمترین پارامترهای موثر بر کیفیت آب شناسایی می‌شوند (Khouri & Al-Mufti 2022). این تکنیک شامل تجزیه و تحلیل تعداد زیادی از پارامترهای کیفیت آب و گروه‌بندی آنها به تعداد کمتری از عوامل است که اکثریت واریانس داده‌ها را توضیح می‌دهد (Shareef., 2019). این تجزیه و تحلیل عاملی در مطالعات مختلف برای ارزیابی کیفیت آب در بدنه‌های آبی مختلف مانند رودخانه‌ها، تالاب‌ها و مخازن استفاده شده است (Ismail & Robescu., 2019, Koryakov, Makar et al. 2023). برای طبقه‌بندی مناطق آلوده، شناسایی منابع آلودگی و نظارت بر تغییرات بلندمدت کیفیت آب استفاده شده است. این روش ثابت کرده است که در ارائه بینش‌های ارزشمند در مورد ساختار شیمیایی بدنه‌های آبی و کمک به مدیریت کیفیت آب موثر است. با تجزیه و تحلیل دقیق اهمیت نسبی ویژگی‌های ورودی و در نظر گرفتن عوامل زمینه‌ای که بر کیفیت آب تأثیر می‌گذارد، محققان می‌توانند درک عمیق‌تری از روابط پیچیده بین متغیرها به دست آورند و مدل‌های مؤثرتری برای پیش‌بینی و مدیریت مسائل کیفیت آب ایجاد کنند (Zhu et al., 2022). در این پژوهش، با استفاده از روش‌های کاهش ابعاد مانند تحلیل عاملی (FA)، پارامترهای کیفیت آب ورودی به شاخص کیفیت آب ایران برای رودخانه مارون مدیریت گردند. با کمک این روش، اطلاعات پیچیده و چند متغیره مربوط به کیفیت آب به یک فضای کم‌ابعادی تبدیل می‌گردد، که در آن متغیرهای مهم و اصلی که بیشترین تاثیر را بر کیفیت آب دارند، به خوبی برجسته خواهند شد. سپس با استفاده از این متغیرهای کم‌ابعادی، الگوریتم‌های یادگیری ماشین به کار گرفته خواهند شد تا بتوانند بهترین پیش‌بینی‌ها را برای شاخص کیفیت آب ارائه دهند. استفاده از الگوریتم‌های یادگیری ماشین با توجه به توانایی آن‌ها در مدل‌سازی روابط پیچیده و غیرخطی، این امکان را فراهم می‌آورند تا الگوهای پنهان در داده‌های کیفیت آب را کشف و پیش‌بینی کنند. نوآوری این پژوهش در این است که با استفاده از ترکیب روش‌های کاهش ابعاد و الگوریتم‌های یادگیری ماشین، امکان مدیریت و پیش‌بینی کیفیت آب ساده‌تر خواهد شد.

مواد و روش‌ها

منطقه موردمطالعه

حوزه آبخیز رودخانه مارون یکی از زیر حوزه‌های رودخانه مارون- جراحی است که در جنوب غربی ایران قرار دارد و در محدوده جغرافیایی ′۰۵-°۵۰ الی ′۱۱-°۵۱ طول شرقی و ′۳۹-°۳۰ الی ′۲۱-°۳۱ عرض جغرافیایی شمالی واقع شده است. رودخانه مارون یکی از آبراه‌های مهم استان خوزستان بوده است که از ارتفاعات زاگرس سر چشمه می‌گیرد و اراضی جنوب شرق استان را مشروب می‌نماید. رودخانه مارون دارای آب دائمی و رژیم بارانی و برفی است به این معنی که بخش عمده ریزش‌های حوزه به صورت باران است. زمین‌شناسی حوزه آبریز رودخانه از سازندهای رخنمون یافته متشکل از ماسه‌سنگ و کنگرمیلا و تناوبی از مارهای رنگی و آهنگ‌های سیلتی می‌باشد. طی گزارش انجام‌شده وجود کارخانه‌های سنگ‌شکن در بستر رودخانه مارون خسارات جبران‌ناپذیری به محیط‌زیست وارد کرده است. ۱۵ واحد صنعتی و خدماتی، یک شهر و تعدادی روستا در حاشیه رودخانه مارون قرار دارد که آلاینده‌های صنعتی و فاضلاب‌های شهری و روستایی را روانه آن می‌کنند. علاوه بر این در قسمت پایین‌دست رودخانه به خصوص از ایدنک به بعد و تا قبل از ورود به دشت بهبهان به واسطه عبور از لایه‌های گچی و نمکی و پس از آن به علت گرما و تبخیر زیاد کیفیت نامطلوبی پیدا می‌کند. در انتهای رودخانه مارون قبل از پیوستن به رودخانه جراحی انواع آلودگی‌های حفاری نفتی پساب‌های کشاورزی و صنعتی به آن وارد می‌شوند.

داده‌ها و پیش‌پردازش آن‌ها

به منظور محاسبه شاخص کیفیت آب ایران ($IRWQI_{SC}$)، داده‌های کیفیت آب همچون دما (T)، نیاز اکسیژن‌خواهی بیوشیمیایی (BOD)، اکسیژن محلول (DO)، کلیفرم کل (T-COLI)، نیترات (NO_3)، فسفات (PO_4)، اسیدیته (pH) و کل مواد جامد محلول (TDS) به صورت ماهانه برای ۱۰ سال آبی (۱۴۰۰-۱۳۹۰) و ایستگاه‌های تنگ تکاب، ایدنک و چم نظام از سازمان آب و برق خوزستان دریافت شد. به منظور بررسی آمار توصیفی داده‌ها مقادیر میانگین، انحراف معیار، مینیمم، ماکزیمم آن‌ها تعیین گردید. توزیع داده‌ها

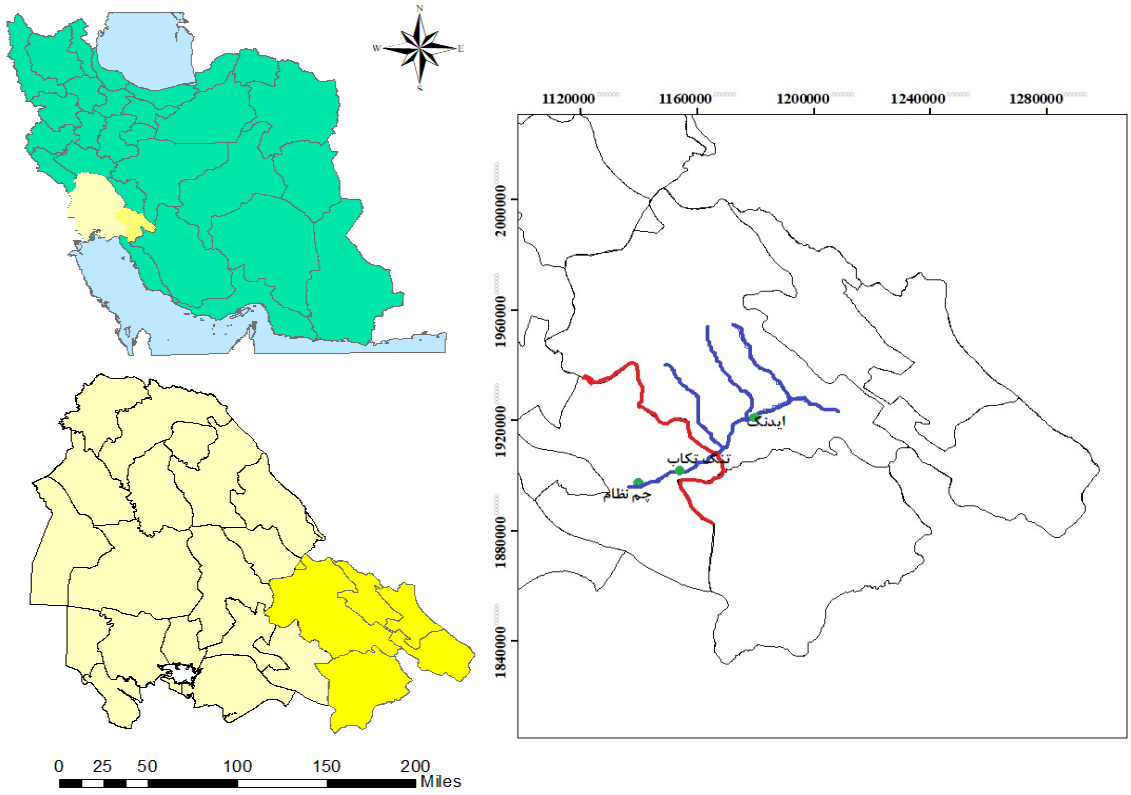
آزمون شاپیرو- ویلک، شاخص‌های کشیدگی و چولگی تعیین گردید.

قابل ذکر است که برای تسهیل در مقایسه داده‌ها و کاهش زمان محاسباتی، تمام داده‌ها با استفاده از معادله (۱) بین صفر و یک نرمال‌سازی شدند

رابطه ۱)
$$x_i^* = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$$

در اینجا X_i^* مقدار نرمال‌سازی‌شده x_i، x_{max} و x_{min}، به ترتیب مقادیر حداکثر و حداقل x_i هستند.

در این تحقیق از دو مجموعه داده برای پیش‌بینی شاخص کیفیت آب ایران استفاده شده است. در مرحله اول از ۸ پارامترهای ذکرشده در بالا به‌عنوان متغیر ورودی استفاده شد. در مرحله بعد داده‌های برجسته با کمک روش‌های کاهش ابعاد استفاده شدند.



شکل ۱- تصویری از منطقه موردمطالعه به همراه ایستگاه‌های هیدرومتری در تحقیق حاضر

کاهش ابعاد داده‌های ورودی با استفاده از روش تحلیل عاملی

تحلیل عاملی یک روش آماری است که با شناسایی الگوهای مشترک بین متغیرها، آن‌ها را به تعداد کمتری از متغیرهای جدید (عامل‌ها) تبدیل می‌کند. این امر می‌تواند به ساده‌تر شدن تحلیل داده‌ها و تفسیر بهتر نتایج کمک کند. آزمون بارتلت، که به عنوان آزمون کرویت بارتلت نیز شناخته می‌شود، یک آزمون آماری است که برای بررسی کروی بودن ماتریس همبستگی در تحلیل عاملی (FA) به کار می رود. به عبارت دیگر، این آزمون فرض می‌کند که همبستگی بین متغیرها در ماتریس همبستگی ضعیف است و ساختار عاملی قابل استخراج است. اگر مقدار p-value آزمون کمتر از سطح معناداری (معمولاً ۰/۰۵) باشد، فرض صفر رد می‌شود و می‌توان نتیجه گرفت که ماتریس همبستگی کروی نیست و ساختار عاملی قابل استخراج است. تحلیل عاملی (Factor Analysis) یکی از روش‌های آماری محبوب در علم داده است که برای کاهش ابعاد مجموعه داده‌ها و خلاصه‌سازی اطلاعات به کار می‌رود. این روش با شناسایی الگوهای پنهان در داده‌ها و دسته‌بندی متغیرهای مرتبط در قالب عوامل جدید، به درک بهتر ساختار داده‌ها و تفسیر آسان‌تر نتایج کمک می‌کند (Stojković., et al 2013). برای انجام آن در تحقیق حاضر ابتدا میزان هم‌خطی بین متغیرهای مستقل با استفاده از ضریب تورم واریانس بررسی می‌شود. هم‌خطی بالا می‌تواند بر دقت تحلیل عاملی تاثیر منفی بگذارد. برای بررسی وجود همبستگی بین متغیرها از آزمون بارتلت استفاده می‌شود. این آزمون فرض می‌کند که ماتریس همبستگی بین متغیرها هویت است. در صورت رد فرضیه صفر،

می‌توان نتیجه گرفت که بین متغیرها همبستگی معنی‌داری وجود دارد و انجام تحلیل عاملی مناسب است. آزمون بارتلت به حساسیت نمونه نسبت به توزیع نرمال داده‌ها شناخته شده است. اگر داده‌ها به طور قابل توجهی از توزیع نرمال فاصله داشته باشند، ممکن است آزمون بارتلت نتایج نادرستی ارائه دهد. ماتریس همبستگی میزان ارتباط بین هر جفت متغیر را نشان می‌دهد. این ماتریس در تحلیل عاملی برای بررسی قدرت همبستگی بین متغیرها و شناسایی متغیرهایی که بیشترین ارتباط را با یکدیگر دارند، استفاده می‌شود. در این مطالعه، از ماتریس همبستگی برای تایید وجود همبستگی قوی بین برخی از متغیرها و دسته‌بندی آن‌ها در قالب عوامل جدید استفاده شد. روش‌های مختلفی برای استخراج عوامل در تحلیل عاملی وجود دارند، از جمله حداکثر واریانس (Maximum Variance)، محورهای اصلی (Principal Components) و چرخش واریانس (Varimax Rotation). در این مطالعه، از روش حداکثر واریانس برای استخراج عوامل استفاده شد. این روش به دنبال یافتن عواملی است که بیشترین واریانس را در مجموعه داده‌ها تبیین می‌کنند. پس از استخراج عوامل، بار عاملی هر متغیر بر روی هر عامل محاسبه می‌شود. بار عاملی نشان‌دهنده میزان ارتباط هر متغیر با یک عامل خاص است. بر اساس بار عاملی، می‌توان هر عامل را بر اساس متغیرهایی که بیشترین ارتباط را با آن دارند، تفسیر و نام‌گذاری کرد.

پیش‌بینی شاخص کیفیت آب ایران با استفاده از الگوریتم‌های یادگیری ماشین

برای پیش‌بینی شاخص کیفیت آب ایران از الگوریتم‌های مختلف یادگیری ماشین همچون Linear Regression، Random Forest، Extra Trees و Light Gradient Boosting Machine استفاده گردید. استفاده از چندین الگوریتم یادگیری ماشین به منظور بهبود دقت پیش‌بینی و انتخاب بهترین مدل انجام شده است و این رویکرد باعث می‌شود که تمامی جوانب مختلف داده‌ها به خوبی مورد بررسی قرار گیرند. Linear Regression یک مدل آماری ساده است که برای مدل‌سازی رابطه خطی بین یک متغیر وابسته (شاخص کیفیت آب) و یک یا چند متغیر مستقل (پارامترهای فیزیکی و شیمیایی آب) استفاده می‌شود. انتخاب ضرایب رگرسیون با استفاده از روش حداقل مربعات معمولی (Ordinary Least Squares) تخمین زده می‌شوند. Random Forest مجموعه‌ای از درختان تصمیم‌گیری تصادفی است که برای پیش‌بینی‌های طبقه‌بندی و رگرسیون به کار می‌رود. این الگوریتم با ترکیب پیش‌بینی‌های درختان مختلف و تنظیم تعداد درختان، عمق درختان دقت و ثبات مدل را افزایش می‌دهد. Extra Trees، درختان اضافی مشابه جنگل تصادفی هستند، با این تفاوت که در هر درخت از تمام ویژگی‌ها به‌جای زیرمجموعه‌ای تصادفی از آن‌ها استفاده می‌شود .این الگوریتم می‌تواند برای شناسایی ویژگی‌های مهم در مجموعه داده مفید باشد. Light Gradient Boosting Machine یک الگوریتم یادگیری ماشین قدرتمند است که از افزایش گرادیان برای یادگیری توابع غیرخطی پیچیده استفاده می‌کند. این الگوریتم به دلیل سرعت و دقت بالا شناخته شده است. در این روش با تنظیم درختان، نرخ یادگیری، عمق درختان، تعداد برگ‌ها در هر درخت کارایی الگوریتم تاثیر می‌پذیرد.

در مرحله اول پیش‌بینی شاخص کیفیت ایران در منطقه موردمطالعه، از داده‌های دما، نیاز اکسیژن‌خواهی بیوشیمیایی، اکسیژن محلول، کلیفرم کل، نیترات، فسفات، اسیدیته و کل مواد جامد محلول و در مرحله دوم تنها از داده‌های با اهمیت‌تری که با استفاده از روش تحلیل عاملی استخراج شده بودند به عنوان ویژگی یا ورودی به الگوریتم‌های هوش مصنوعی استفاده گردید. داده‌های مورداستفاده در هر مرحله به دو گروه تقسیم شد به طوری که ۸۰٪ از داده‌های موجود برای آموزش و ۲۰٪ برای آزمایش استفاده گردید. سپس عملکرد الگوریتم‌های موردمطالعه در پیش‌بینی شاخص کیفیت آب با استفاده از معیارهای متداول ارزیابی همچون ریشه میانگین مربعات خطا (RMSE) و ضریب تعیین (R^2) تعیین گردید (Hafshejani et al., 2024).

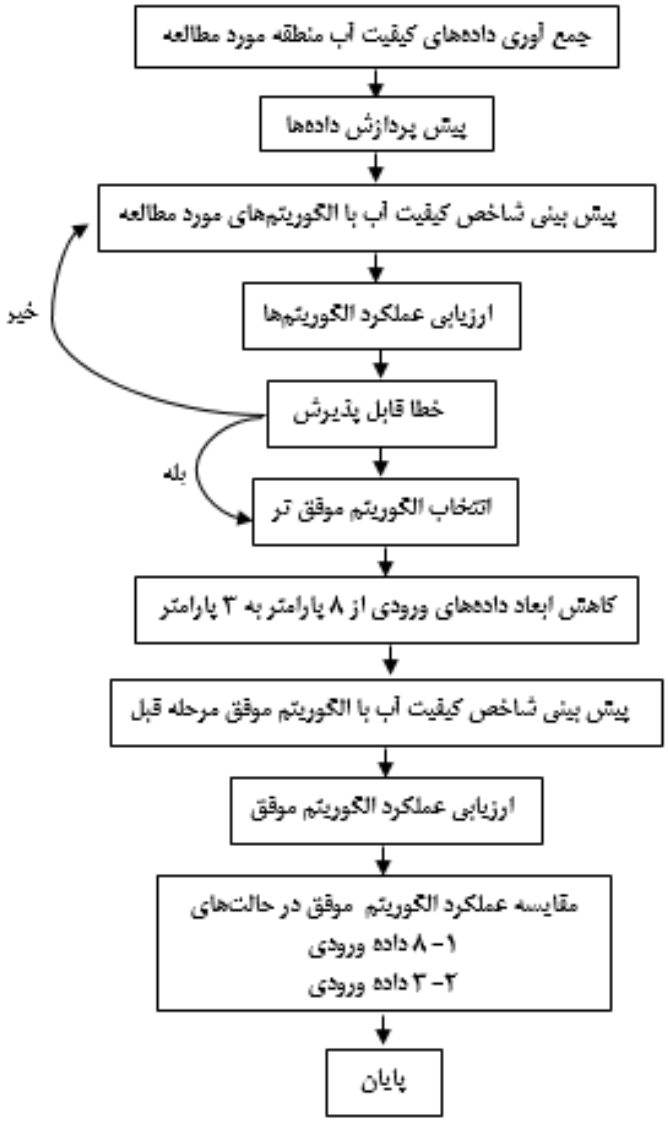
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i^{exp} - Y_i^{pred})^2} \quad \text{رابطه ۲)}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|Y_i^{exp} - Y_i^{pred}\right|}{n} \quad \text{رابطه ۳)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(Y_I^{exp} - Y_{ave}^{exp}\right)^2} \quad \text{رابطه ۴)}$$

در فرمول‌های بالا، Y_i^{pred}: مقدار iام شاخص پیش‌بینی شده توسط مدل، Y_i^{exp}: مقدار iام شاخص محاسبه‌شده: Y_{ave}^{exp}، مقدار متوسط شاخص محاسبه‌شده و n: تعداد داده‌ها است.

ریشه میانگین مربعات خطا نشان‌دهنده‌ی میزان تطابق بین مقادیر پیش‌بینی شده توسط مدل و مقادیر واقعی است. مقدار کمتر آن نشان‌دهنده‌ی تطابق بهتر مدل با داده‌ها است. ضریب تعیین نسبت واریانس در متغیر وابسته را که توسط متغیرهای مستقل توضیح می‌شود را اندازه‌گیری می‌کند. مقدار بالاتر آن نشان‌دهنده‌ی تطابق بهتر مدل با داده‌ها است.



شکل ۲- مراحل انجام تحقیق حاضر

نتایج و بحث

نتایج پیش‌پردازش داده‌ها

اطلاعات مربوط به ۸ پارامتر مختلف شاخص کیفیت آب ایران (TDS، T، BOD، DO، T- COLI، NO^{-}_{3}، PO_4، pH) در جدول (۱) نشان داده شده است.

جدول ۱- تجزیه آماری پارامترهای مختلف شاخص کیفیت آب ایران

پارامتر	دما	نیاز اکسیژن خواهی بیوشیمیایی	اکسیژن محلول	کلیفرم کل	نیترات	فسفات	اسیدی ته	کل مواد جامد محلول
میانگین	۱۹/۵۸	۲/۵۸	۸/۰۴	۲۵۳۱۷	۶/۷۰	۰/۰۲	۷/۷۸	۱۴۴۹/۶۱
حداقل	۱۰	۰/۷	۴/۳	۲۳۰	۱/۵۸	۰	۷/۴	۱۶۸/۷
حداکثر	۳۰	۴/۸۲	۱۰/۶	۱۱۰۰۰۰	۱۷/۳۱	۰/۱۲	۸/۱	۲۰۰۲

انحراف معیار	۴/۴۰	۰/۸۳	۱/۰۹۸	۳۰۹۶۲/۲۰	۲/۶۰	۰/۰۲	۰/۱۷	۲۸۹/۰۳

این پارامترها حاوی اطلاعات مهمی برای ارزیابی کیفیت آب و سلامت آن هستند. انحراف معیار بالا برایBOD، T-COLI، NO_3، PO_4، pH نشان‌دهنده تنوع زیاد در مقادیر این پارامترها در بین نمونه‌ها است. این تنوع بالا می‌تواند نشان‌دهنده ناپایداری کیفیت آب و یا وجود منابع آلودگی نقطه‌ای باشد (Giao et al., 2022). مقادیر حداقل و حداکثر نشان‌دهنده محدوده تغییرات پارامترها در بین نمونه‌ها هستند. برای مثال، وجود مقادیر بالای BOD و T-COLI در برخی نمونه‌ها نشان‌دهنده آلودگی نقطه‌ای در برخی از منابع آب است.

مقدار p-value در آزمون شاپیرو- ویلک نیز معیاری است که نشان‌دهنده تطابق داده‌ها با یک توزیع نرمال است. با توجه به مقدار p-value برای همه پارامترهای کیفیت آب که بیشتر از سطح معناداری ۰/۰۵ می‌باشند، فرض صفر تائید و بنابراین مشخص گردید داده‌ها از یک توزیع نرمال پیروی می‌کنند.

جدول ۲- نتایج آزمون شاپیرو- ویلک برای تعیین نرمال بودن داده‌های تحقیق حاضر

پارامتر	TDS	pH	PO_4	NO_3	T-COLI	DO	BOD	T
مقدار p-value	۰/۹۰	۰/۹۵	۰/۸۳	۰/۸۸	۰/۷۱	۰/۹۹	۰/۹۹	۰/۹۸

تحلیل عاملی و کاهش ابعاد داده‌ها

مقدار p-value آزمون بارتلت در تحقیق حاضر نزدیک صفر بود (۰/۰۰۰۰۸) و می‌توان نتیجه گرفت که داده‌ها برای تحلیل عاملی و کاهش ابعاد مناسب هستند (Patil et al., 2020). نتایج ماتریس همبستگی که نشان‌دهنده ارتباط بین هر جفت متغیر است در جدول (۳) ارائه گردید.

همان‌طور که در جدول ۳ مشاهده می‌شود ارتباط منفی بین BOD و DO با ضریب همبستگی ۰/۳۵- برقرار است. این به این معنی است که با افزایش مقدارBOD ، مقدار DO به‌طور معنی‌داری کاهش می‌یابد. کاهش DO با افزایشBOD ، پدیده‌ای شناخته‌شده در سیستم‌های آبی است. BOD، مخفف تقاضای نیاز اکسیژن خواهی بیوشیمیایی است و نشان‌دهنده مقدار اکسیژن موردنیاز میکروارگانیسم‌ها برای تجزیه مواد آلی موجود در آب است. هنگامی که BOD بالا باشد، به این معنی است که مواد آلی آلاینده زیادی در آب وجود دارد که میکروارگانیسم‌ها برای تجزیه آن‌ها به اکسیژن نیاز دارند. در نتیجه، با افزایشBOD ، اکسیژن محلول در آب (DO)که برای تنفس موجودات زنده ضروری است، به طور قابل‌توجهی کاهش می‌یابد. بنابراین، مشاهده ارتباط منفی بین BOD و DOدر ماتریس همبستگی، نشان‌دهنده وجود آلودگی در آب و اختلال در سلامت اکوسیستم آبی است. این امر ضرورت اتخاذ اقدامات لازم برای کنترل BOD و حفظ DO در سطوح مناسب را برای حفظ سلامت و پایداری اکوسیستم‌های آبی آشکار می‌کند.

مقدار همبستگی بین (NO_3) و (IRWQIsc) برابر با ۰/۵۶- است. این مقدار همبستگی نشان می‌دهد که با افزایش غلظت نیترات در آب، شاخص کیفیت آب کاهش می‌یابد. نیترات معمولاً از فعالیت‌های زیستی و آلاینده‌های آنتروپوژنیک مانند کودها و فاضلاب‌های کشاورزی به آب اضافه می‌شوند. افزایش نیترات می‌تواند به افزایش رشد گیاهان آبزی، کاهش تنوع زیستی در آب و در نتیجه کاهش کیفیت آب منجر شود.

مقدار همبستگی بین (TDS) و (IRWQIsc) برابر با ۰/۴۸- است. این نشانگر این است که با افزایش مقدار TDS، شاخص کیفیت آب کاهش می‌یابد. مواد جامد حل‌شده در آب می‌توانند از منابع مختلفی مانند معادن، کشاورزی، و فعالیت‌های صنعتی به آب اضافه شوند. افزایش مقدار TDS ممکن است نشانگر آلودگی آب با مواد شیمیایی و معدنی باشد که می‌تواند بر کیفیت آب تأثیر داشته باشد.

جدول ۳- ماتریس همبستگی پارامترهای مختلف شاخص کیفیت آب ایران در تحقیق حاضر

	T	BOD	DO	T-COLI	NO_3	PO_4	pH	TDS	$IRWQI_{sc}$
T	۱	۰/۰۲	۰/۲۶	-۰/۱۳	۰/۳۰	۰	۰/۱۱	۰/۰۶	۰
BOD	۰/۰۲	۱	-۰/۳۵	۰	۰/۰۵	۰/۰۳	۰/۰۱	۰/۰۹	-۰/۰۵
DO	۰/۲۶	-۰/۳۵	۱	۰/۰۸	۰/۰۵	۰/۰۹	۰/۰۷	-۰/۰۵	۰/۱۵

T-COLI	-۰/۱۳	۰	۰/۰۸	۱	-۰/۰۷	-۰/۰۱	۰/۱	۰/۱	-۰/۳
NO_3	۰/۳	۰/۰۵	۰/۰۵	-۰/۰۷	۱	۰/۰۵	۰/۱	۰/۳۲	-۰/۵۶
PO_4	۰	۰/۰۳	۰/۰۹	-۰/۰۱	۰/۰۵	۱	-۰/۰۲	-۰/۰۲	-۰/۰۸
pH	۰/۱۱	۰/۰۱	۰/۰۷	۰/۱	۰/۱	-۰/۰۲	۱	-۰/۰۲	۰/۰۲
TDS	۰/۰۶	۰/۰۹	-۰/۰۵	۰/۱	۰/۳۲	-۰/۰۲	-۰/۰۲	۱	-۰/۴۸
IRWQIsc	۰	-۰/۰۵	۰/۱۵	-۰/۳	-۰/۵۶	-۰/۰۸	۰/۰۲	-۰/۴۸	۱

مقادیر ضریب تورم واریانس برای پارامترهای کیفیت آب مورداستفاده در تحقیق حاضر (جدول ۴) نشان داد که متغیرهای کلیفرم کل و فسفات هم‌خطی کمی با سایر متغیرهای مستقل دارند.

جدول ٤- ضریب تورم واریانس پارامترهای مختلف شاخص کیفیت آب ایران در تحقیق حاضر

پارامتر	مقدار
دما	۷/۸۲
نیاز اکسیژن خواهی بیوشیمیایی	۷/۱۷
اکسیژن محلول	۱۸/۶۳
کلیفرم کل	۱/۷۹
نیترات	۷/۳۰
فسفات	۲/۶۸
اسیدیته	۵/۹۰
کل مواد جامد محلول	۱۸/۲۸

اطمینان از عدم وجود هم‌خطی مضر بین متغیرها ضروری است، زیرا هم‌خطی می‌تواند بر دقت نتایج تحلیل عاملی تاثیر بگذارد. مقدار ضریب تورم واریانس برابر ۱، نشان‌دهنده عدم وجود چند خطی است، در حالی که مقادیر بالاتر از ۱ نشان‌دهنده افزایش چند خطی بودن است. نمره ضریب تورم واریانس بالای ۵ یا ۱۰ معمولاً بالا در نظر گرفته می‌شود و نشان‌دهنده چند خطی بودن شدید است (Kyriazos & Poga, 2023). مقدار ضریب تورم واریانس متغیرهای دما، نیاز اکسیژن خواهی بیوشیمیایی، اکسیژن محلول، نیترات، اسیدیته و کل مواد جامد محلول بیشتر از ۵ باشد، که نشان‌دهنده هم‌خطی بین آن‌ها و سایر متغیرها است.

تفسیر نام و مفهوم عوامل

مقادیر بارهای عاملی استخراج‌شده با روش حداکثر واریانس برای هر متغیر در جدول ۵ نشان داده شده است. این مقادیر نشان‌دهنده میزان سهم هر یک از پارامترهای اصلی در عوامل استخراج‌شده هستند. پارامترهایی با بار نزدیک به ۱ (مثبت یا منفی) نشان‌دهنده همبستگی قوی بین پارامتر و عامل است. مقادیر نزدیک به ۰ نشان‌دهنده همبستگی ضعیف‌تر است. رابطه معکوس بین دما و اکسیژن محلول در جدول ۵، بیانگر این نکته است که آب گرم‌تر اکسیژن محلول کمتری را در خود نگه می‌دارد. ارتباط بالقوه بین فعالیت بیولوژیکی و آلودگی باکتریایی در جدول ۵ نشان‌دهنده سطوح بالاتر نیاز اکسیژن خواهی بیوشیمیایی با افزایش رشد میکروبی است. ارتباط بین سطوح نیترات و فسفات احتمالاً منابع آلودگی کشاورزی یا صنعتی را نشان می‌دهد. رابطه بین اسیدیته و کل مواد جامد محلول نشان می‌دهد که شرایط اسیدی ممکن است بر حلالیت جامدات محلول تأثیر بگذارد.

جدول ۵- بارهای عاملی پارامترهای مختلف شاخص کیفیت آب ایران در تحقیق حاضر

پارامتر	عامل ۱	عامل ۲	عامل ۳
دما	۰/۲۴	۰/۳۸	۰/۱۵
نیاز اکسیژن خواهی بیوشیمیایی	۰/۴۱	۰/۰۳	-۰/۰۲
اکسیژن محلول	-۰/۹۸	-۰/۱۸	۰/۰۵
کلیفرم کل	۰/۰۸	-۰/۲۴	۰/۱۴
نیترات	۰/۱۹	۰/۰۷۶	۰/۱۴
فسفات	۰/۱۱	۰/۰۷	۰/۰۱

اسیدیته	-۰/۰۱	۰/۰۳	۰/۸۵
کل مواد جامد محلول	۰/۰۸	۰/۳۷	-۰/۰۵

بر اساس بارهای عاملی ارائه‌شده، می‌توان نام‌ها و تفاسیر مفهومی زیر را برای هر عامل پیشنهاد کرد:

عامل ۱ (آلودگی آلی): این عامل با بارهای مثبت بر روی نیاز اکسیژن خواهی بیوشیمیایی، نیترات و فسفات مشخص می‌شود که نشان‌دهنده ارتباط با مواد آلی و سطوح مواد مغذی است. بار منفی روی DO نشان می‌دهد که تجزیه مواد آلی اکسیژن محلول را مصرف می‌کند. بار مثبت بر روی دمای آب بیانگر این است که وجود آلاینده‌های آلی باعث بالا رفتن دما می‌شود. رشد بیش از حد جلبک‌ها و گیاهان آبزی، که توسط آلودگی مواد مغذی مانند نیترات‌ها و فسفات‌ها تغذیه می‌شود، منجر به کاهش سطح اکسیژن محلول (Huang et al., 2022) و افزایش دمای آب می‌شود (Varghese & Gunasundari., 2024). علاوه بر این، ذرات آلی که باعث افزایش کدورت در آب می‌شوند، منجر به جذب بیشتر انرژی خورشیدی و افزایش بیشتر دمای آب می‌شود.

عامل ۲ (کدورت آب یا کل جامدات محلول): این عامل دارای بارهای مثبت بر روی کل مواد جامد محلول و دما است. حلالیت مواد جامد محلول در آب با افزایش دما به طور کلی افزایش می‌یابد. این امر به دلیل افزایش حرکت مولکولی، کاهش کشش سطحی و تغییر در ساختار آب با گرم شدن آن است. درک رابطه بین دمای آب و مواد جامد محلول برای ارزیابی کیفیت آب، طراحی سامانه‌های تصفیه آب و درک فرآیندهای طبیعی مانند رسوب‌گذاری و فرسایش مهم است (Jakubowicz et al., 2022; Adjovu et al., 2023). از طرفی بالا رفتن دما می‌تواند بر حلالیت اکسیژن در آب تأثیر منفی بگذارد و به دنبال آن بر دسترسی اکسیژن محلول برای باکتری‌های هوازی مانند کلیفرم‌ها تأثیر منفی بگذارد. سطح اکسیژن محلول در آب نقش مهمی در حمایت از باکتری‌های هوازی مانند کلیفرم‌ها ایفا می‌کند. سطوح بالاتر اکسیژن محلول می‌تواند بقا و رشد باکتری‌های کلیفرم را در بدنه‌های آبی افزایش دهد.

عامل ۳ (اسیدیته): این عامل بار مثبت بالایی بر روی pH و بعد از آن دما دارد. افزایش دما باعث افزایش pH آب می‌شود. این امر به دلیل کاهش انحلال گازهای اسیدی مانند دی اکسیدکربن در آب با افزایش دما است. در ارتباط با رابطه pH و نیترات، می‌توان به این مورد اشاره کرد که کاهش pH باعث می‌شود که نیترات به اسید نیتریک تبدیل شود و مقدار آن هم کاهش یابد. علاوه بر این در pH پایین، رشد باکتری‌های کلیفرم محدود می‌شود.

پیش‌بینی شاخص کیفیت آب ایران با استفاده از الگوریتم‌های یادگیری ماشین

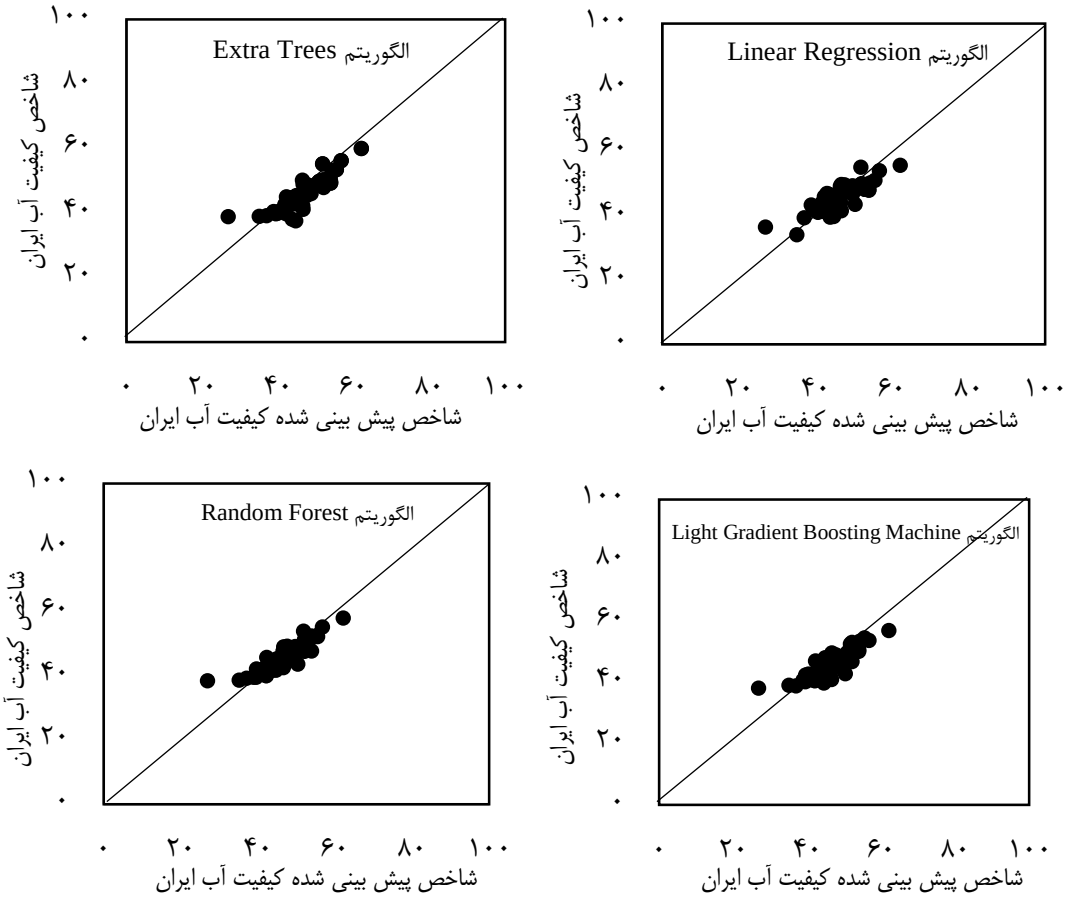
نتایج پیش‌بینی شاخص کیفیت آب با استفاده از ۸ پارامتر موردمطالعه به عنوان ورودی در شکل ۳ و جدول ۶ نشان داده شده است. در میان الگوریتم‌های مختلف موردبررسی، رگرسیون خطی کمترین کارایی را در پیش‌بینی شاخص کیفیت آب ایران داشت. این الگوریتم با R^2 برابر با ۰/۷۰، که پایین‌ترین مقدار در بین الگوریتم‌ها است، و همچنین با مقادیر بالای RMSE (۳/۰۶) و MAE (۲/۳۷)، نشان داد که در مدل‌سازی رابطه بین پارامتر کیفیت آب ورودی و شاخص کیفیت آب ناموفق بوده است. علت عدم کارایی رگرسیون خطی در این مورد، سادگی ذاتی این الگوریتم است. رگرسیون خطی از یک خط مستقیم برای پیش‌بینی مقادیر متغیر وابسته (شاخص کیفیت آب) بر اساس مقادیر متغیرهای مستقل (پارامترهای کیفیت آب ورودی) استفاده می‌کند. این در حالی است که رابطه بین این دو متغیر به‌احتمال زیاد غیرخطی و پیچیده‌تر از یک خط مستقیم است. به همین دلیل، رگرسیون خطی قادر به درک و مدل‌سازی دقیق این رابطه نبوده است (Belzak & Bauer, 2019).

بالاترین دقت پیش‌بینی با R^2 (۰/۷۸)، RMSE (۲/۶۵) و MAE (۱/۹۰)، مربوط به الگوریتم جنگل تصادفی بود. میزان اهمیت ویژگی‌های ورودی به هر الگوریتم در شکل ۴ نشان داده شده است. همان‌طور که در شکل نشان داده شده است در ۳ الگوریتم موردمطالعه، نیترات و اسیدیته به ترتیب پر و کم اهمیت‌ترین پارامترهای ورودی به آن‌ها هستند. بدیهی است که پارامترهای با اهمیت نقش مهمی در توانایی الگوریتم‌ها در پیش‌بینی متغیر هدف دارد (Khaire & Dhanalakshmi, 2022). این نتیجه که نیترات مهمترین پارامتر ورودی و اسیدیته دارای اهمیت کمتری برای سه الگوریتم مورد مطالعه است با یافته‌های تحلیل عاملی مطابقت دارد. ارتباط قوی نیترات با آلودگی آلی و ارتباط ضعیف‌تر اسیدیته با شفافیت آب این تفسیر را تأیید می‌کند. با این حال، در نظر گرفتن محدودیت‌های تحلیل عاملی و مدل‌های یادگیری ماشین و ادغام دانش برای درک جامع عوامل مؤثر بر کیفیت آب بسیار مهم است.

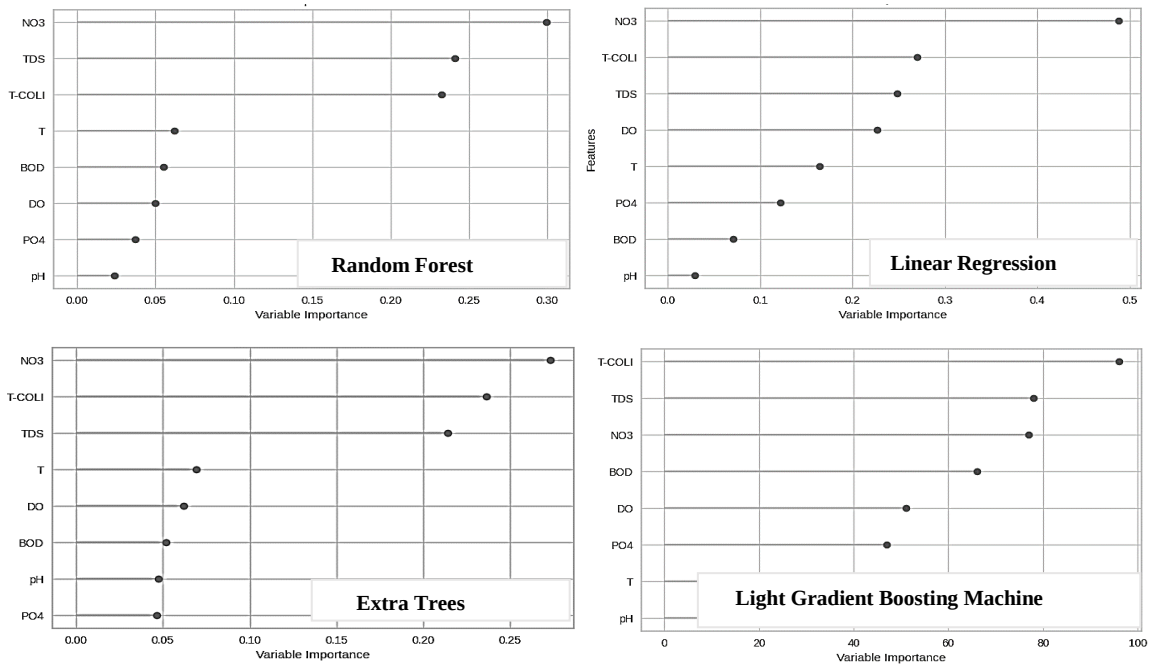
جدول ٦- معیارهای ارزیابی الگوریتم‌های موردمطالعه در تحقیق حاضر (۸ پارامتر ورودی)

MAE	RMSE	R^2	الگوریتم
۱/۹۰	۲/۶۵	۰/۷۸	Random Forest
۲/۰۹	۲/۸۰	۰/۷۵	Light Gradient Boosting Machine
۲/۱۳	۲/۹۴	۰/۷۲	Extra Trees
۲/۳۷	۳/۰۶	۰/۷۰	Linear Regression

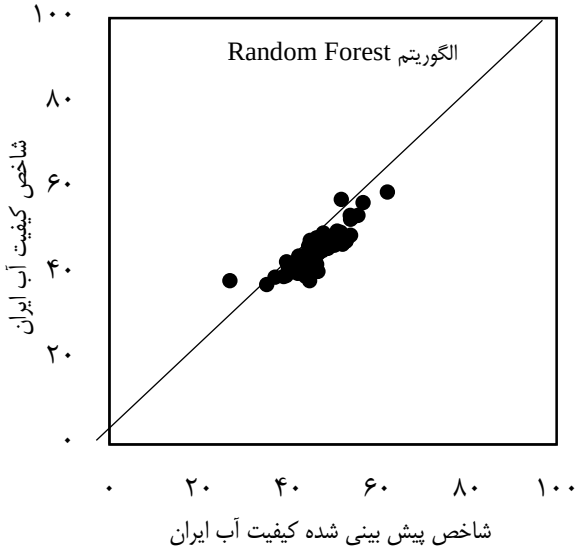
از آنجایی که یکی از اهداف مدل‌های هوش مصنوعی کاهش زمان عملیاتی و هزینه مطالعات آتی است، نتایج الگوریتم جنگل تصادفی (مدل موفق مرحله قبل) در پیش‌بینی شاخص کیفیت آب ایران تنها در شرایطی که پارامترهای ورودی از ۸ ویژگی کیفیت آب به ۳ ویژگی استخراج‌شده از نتایج کاهش ابعاد تغییر یافتند در شکل ۵ و جدول ۶ نشان داده شده است. مقایسه‌ی مقادیر MAE، RMSE و R^2 برای مدل‌های جنگل تصادفی با ۸ ورودی و ۳ ورودی نشان داد که نتایج شبیه‌سازی توسط این الگوریتم با ۳ ورودی نیز تقریباً مشابه ۸ ورودی است و بنابراین می‌تواند جایگزین آن شود.



شکل ۳- نتایج پیش‌بینی شاخص کیفیت آب توسط الگوریتم‌های موردمطالعه با ۸ پارامتر ورودی



شکل ٤- درجه اهمیت پارامترهای ورودی به الگوریتم‌های موردمطالعه با ۸ پارامتر ورودی



شکل ٥- نتایج پیش‌بینی شاخص کیفیت آب توسط الگوریتم بهینه Random Forest با ۳ پارامتر ورودی

جدول ٦- معیارهای ارزیابی الگوریتم بهینه در تحقیق حاضر (۳ پارامتر ورودی)

MAE	RMSE	R^2	الگوریتم
۲/۰۹	۲/۸۶	۰/۷۴	Random Forest

نتیجه‌گیری

یافته‌های این پژوهش نشان داد که از میان الگوریتم‌های مورداستفاده در تحقیق حاضر برای پیش‌بینی شاخص کیفیت آب ایران در رودخانه مارون با ۸ پارامتر ورودی، الگوریتم جنگل تصادفی با R^2 (۰/۷۸) و RMSE (۲/۶۵) بهترین عملکرد و الگوریتم رگرسیون خطی با R^2 (۰/۷۰) و RMSE (۳/۰۶) پایین‌ترین درجه موفقیت را داشت. نتایج حاصل از تحلیل اهمیت ویژگی نشان داد که NO_3^-، مهم‌ترین پارامتر در پیش‌بینی شاخص کیفیت آب رودخانه مارون است و بنابراین اندازه‌گیری پیوسته و دقیق آن باید در برنامه‌ریزی منابع آب مورد توجه قرار گیرد. از طرفی pH دارای اهمیت کمتری نسبت به سایر پارامترها است. نتایج تحلیل عاملی نشان داد که سه

پارامتر کلیدی (NO^-_3، TDS و T-COLI) نقش تعیین‌کننده‌ای در پیش‌بینی شاخص کیفیت آب در منطقه موردمطالعه دارند. نتایج شبیه‌سازی توسط الگوریتم جنگل تصادفی با پارامترهای ورودی NO^-_3، TDS و T-COLI نیز تقریبا مشابه (R^2 (۰/۷۴) و RMSE (۲/۸۶)) با نتایج پیش‌بینی با پارامترهای ورودی TDS، T، BOD، DO، T- COLI، NO^-_3، PO_4، pH است و بنابراین از آنجایی که هدف هوش مصنوعی کاهش زمان و هزینه در مطالعات است پیش‌بینی با ۳ پارامتر می‌تواند جایگزین پیش‌بینی با ۸ پارامتر ورودی گردد.

سپاس‌گزاری

بدین‌وسیله از حمایت مالی معاونت پژوهش و فناوری دانشگاه شهید چمران اهواز در قالب پژوهانه (SCU.WE1402.47794) در انجام این تحقیق تشکر و قدردانی می‌گردد.

"هیچ‌گونه تعارض منافع بین نویسندگان وجود ندارد"

REFERENCES

Adjovu, G. E., Stephen, H., & Ahmad, S. (2023). A machine learning approach for the estimation of total dissolved solids concentration in Lake Mead using electrical conductivity and temperature. *Water*, *15*(13), 2439.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water*, *11*(11), 2210.

Ali, N., Chen, J., Fu, X., Hussain, W., Ali, M., Iqbal, S. M., Anees, A., Hussain, M., Rashid, M., & Thanh, H. V. (2023). Classification of reservoir quality using unsupervised machine learning and cluster analysis: Example from Kadanwari gas field, SE Pakistan. *Geosystems and Geoenvironment*, *2*(1), 100123.

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2022). Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment*, *8*(2), 2793-2801.

Belzak, W. C., & Bauer, D. J. (2019). Interaction effects may actually be nonlinear effects in disguise: A review of the problem and potential solutions. *Addictive behaviors*, *94*, 99-108.

Chen, B., Mu, X., Chen, P., Wang, B., Choi, J., Park, H., Xu, S., Wu, Y., & Yang, H. (2021). Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data. *Ecological Indicators*, *133*, 108434.

Chen, P., Wang, B., Wu, Y., Wang, Q., Huang, Z., & Wang, C. (2023). Urban River water quality monitoring based on self-optimizing machine learning method using multi-source remote sensing data. *Ecological Indicators*, *146*, 109750.

Deng, T., Chau, K.-W., & Duan, H.-F. (2021). Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management*, *284*, 112051.

Divband Hafshejani, L., Naseri, A. A., Moradzadeh, M., Daneshvar, E., & Bhatnagar, A. (2022). Applications of soft computing techniques for prediction of pollutant removal by environmentally friendly adsorbents (case study: the nitrate adsorption on modified hydrochar). *Water Science & Technology*, *86*(5), 1066-1082.

Giao, N. T., Nhien, H. T. H., Anh, P. K., & Thuptimdang, P. (2022). Combination of water quality, pollution indices, and multivariate statistical techniques for evaluating the surface water quality variation in Can Tho City, Vietnam. *Environmental Monitoring and Assessment*, *194*(11), 844.

Hafshejani, L. D., Naseri, A. A., Hooshmand, A., Mohammadi, A. S., & Abbasi, F. (2024). Prediction of nitrate leaching from soil amended with biosolids by machine learning algorithms. *Ain Shams Engineering Journal*, 102783.

Haggerty, R., Sun, J., Yu, H., & Li, Y. (2023). Application of machine learning in groundwater quality modeling-A comprehensive review. *Water Research*, 119745.

Huang, M. V. (2022). Impact of Environmental Factors on the Algae Overgrowth in Pond Water. *Journal of Student Research*, 11(3).

Ismail, A. H., & Robescu, D. (2019). Application of multivariate statistical techniques in water quality assessment of Danube river, Romania. *Environ. Eng. Manag. J*, *18*, 719-726.

Jakubowicz, P., Steliga, T., & Wojtowicz, K. (2022). Analysis of Temperature Influence on Precipitation of Secondary Sediments during Water Injection into an Absorptive Well. *Energies*, *15*(23), 9130.

Jatnika, H., Huda, M., Amelia, R. R., Manuhutu, M. A., Windarto, A. P., Sumantrie, P., & Waluyo, A. (2021, February). Analysis of data mining in the group of water pollution areas using the K-means method in Indonesia. In *Journal of Physics: Conference Series* (Vol. 1783, No. 1, p. 012014). IOP Publishing.

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, *34*(4), 1060-1073.

Khouri, L., & Al-Mufti, M. B. (2022). Assessment of surface water quality using statistical analysis methods: Orontes River (Case study). *Baghdad Science Journal*, *19*(5), 0981-0981.

Koryakov, A., Makar, S., Lukyanets, A., & Moreva, E. (2023). Peculiarities of Statistical Water Quality Assessment in an Industrial Region. *Polish Journal of Environmental Studies*, *32*(1).

Krishnan, S., & Manikandan, R. (2024). Water quality prediction: A data-driven approach exploiting advanced machine learning algorithms with data augmentation. *Journal of Water and Climate Change*.

Kyriazos, T., & Poga, M. (2023). Dealing with multicollinearity in factor analysis: the problem, detections, and solutions. *Open Journal of Statistics*, *13*(3), 404-424.

Li, Y., Mi, W., Ji, L., He, Q., Yang, P., Xie, S., & Bi, Y. (2023). Urbanization and agriculture intensification jointly enlarge the spatial inequality of river water quality. *Science of the Total Environment*, *878*, 162559.

Patil, V. B., Pinto, S. M., Govindaraju, T., Hebbalu, V. S., Bhat, V., & Kannanur, L. N. (2020). Multivariate statistics and water quality index (WQI) approach for geochemical assessment of groundwater quality—a case study of Kanavi Halla Sub-Basin, Belagavi, India. *Environmental Geochemistry and Health*, *42*, 2667-2684.

Schäfer, B., Beck, C., Rhys, H., Soteriou, H., Jennings, P., Beechey, A., & Heppell, C. M. (2022). Machine learning approach towards explaining water quality dynamics in an urbanised river. *Scientific Reports*, *12*(1), 12346.

Shareef, M. A. (2019). Assessment of Tigris River water quality using multivariate statistical techniques. *Tikrit Journal of Engineering Sciences*, *26*(4), 26-31.

Sharma, V., Sharma, M., Pandita, S., Kumar, V., Kour, J., & Sharma, N. (2021). Assessment of water quality using different pollution indices and multivariate statistical techniques. In *Heavy metals in the environment*: 165-178.

Stojković, J., Papić, P., Ćuk, M., & Todorović, M. (2013). Application of factor analysis in identification of dominant hydrogeochemical processes of some nitrogenous groundwater of Serbia. *Geoloski anali Balkanskoga poluostrva*, (74), 57-62.

Varghese, I. S., & Gunasundari, R. (2024). Cubic Grey Relational Luong Attention Bidirectional Long Short-Term Memory based Dissolved Oxygen Prediction in River. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(11s), 387-395.

Watkins, K. (2006). Human Development Report 2006-Beyond scarcity: Power, poverty and the global water crisis. *UNDP Human Development Reports (2006)*.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*.