Water and Soil Management and Modeling

Online ISSN: 2783 - 2546

University of Mohaghegh Ardabili

Evaluating the efficiency of dimensionality reduction methods in improving the accuracy of water quality index modeling in Qizil-Uzen River using machine learning algorithms

Mohammad Taghi Sattari [1*], Kimia Shirini [2], Sahar Javidan [3]

[1] Associate Professor, Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[2] Ph.D. student, Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran
[3] M.Sc. Student, Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran

Abstract

Introduction

Water quality assessment is paramount for various sectors, including environmental planning, public health, and industrial operations. With the increasing importance of ensuring safe water sources, especially for drinking and irrigation purposes, modern methodologies like data mining offer valuable tools for predictive analysis and classification of water quality. Knowledge of water quality is considered one of the most important needs in planning, developing, and protecting water resources. Determining the quality of water for different uses, including irrigation and drinking in different areas of life. The use of modern data mining methods can be beneficial for predicting and classifying the quality of provider water. In the current study, the water quality of the Qizil-Uzen River was evaluated at Qara Gunei stations. In this regard, the drinking water quality index (WQI) using the chemical compounds of glass hardness, alkalinity (PH), electrical conductivity, total dissolved substances, calcium, sodium, magnesium, potassium, chlorine, carbonate, bicarbonate and sulfate in the statistical period of 21 years (2000-2020) was estimated. Water quality assessment is paramount for various sectors, including environmental planning, public health, and industrial operations. With the increasing importance of ensuring safe water sources, especially for drinking and irrigation purposes, modern methodologies like data mining offer valuable tools for predictive analysis and classification of water quality.

Materials and Methods

Due to the relatively large number of variables, principal component analysis and independent component analysis methods were used to reduce dimensions, and then different machine learning algorithms including decision tree, logistic regression, and multi-layer perceptron artificial neural network were used to model the water quality index. By using these methods, the number of parameters needed to calculate the quality index was reduced from 12 to 2. Reducing the dimensions of the data saves the time of sampling, monitoring the samples, and determining the quality of the water and reduces the costs required for modeling to a significant amount. The results showed that among the dimensionality reduction methods, the principal component analysis method can perform better than the independent component analysis method. In the current research, the WQI index was modeled using machine learning algorithms including decision tree, logistic regression, and artificial neural network method. The quality of water in the Qizil-Uzen Qara Gunei river station has been evaluated. Then, to estimate the numerical values of the WQI index, TH, pH, EC, TDS, Ca, Na, Mg, K, Cl, CO3, HCO3, and SO4 parameters of the mentioned station in the statistical period of 21 years (1378-1398) were used. PCA and ICA methods have been used to select different input parameters. Modeling has been done in a Python programming environment. Among the available samples, 75% are considered for training and 25% for testing.

Results and Discussion

In the present research, to model the water quality index in the first stage, different dimensionality reduction methods such as PCA and ICA were used to reduce the time and cost of implementation. In the second stage, machine learning methods such as decision tree, linear regression, and multilayer perceptron were used. In the method used by Tripathi and his colleagues, by using the principal component analysis method, they reduced the number of parameters needed to calculate the quality index from 28 to 9 and calculated the water quality index with the number of 9 parameters. Examining the two methods of PCA and ICA has reduced the dimensions of the problem from 12 dimensions to 2 dimensions. The results show that the PCA method can help us improve performance with little cost and high accuracy. Because of the PCA dimensions. The comparison of the results of the models was done using different numerical and graphical evaluation criteria, including R^2, RMSE, and

Water and Soil Management and Modeling

Online ISSN: 2783 - 2546

University of Mohaghegh Ardabili

modified Wilmot coefficient as numerical criteria and Taylor diagram as graphical criteria. Because the PCA algorithm can help reduce noise in data, feature selection, and generate independent and unrelated features from data. The results show that multi-layer perceptron, decision tree, and logistic regression methods accurately perform the water quality index. In this research, for the first time, using the ICA dimension reduction algorithm, while reducing the dimensions of the problem, the water quality index is predicted with an accuracy of over 90%.

Conclusion

Water quality index modeling holds significant relevance in agricultural practices, where access to clean water is crucial for irrigation and crop growth. Surprisingly, only a limited number of studies have explored variable reduction methods in water quality index modeling, with none incorporating the relatively novel Independent Component Analysis (ICA) method for dimensionality reduction. Thus, the current research fills this gap by employing PCA and ICA techniques to reduce the dimensionality of large datasets in water quality index modeling. By utilizing these advanced methods, the study aims to enhance efficiency and accuracy in assessing water quality, thereby offering valuable insights for agricultural water management. Following dimensionality reduction, the dataset is then subjected to modeling using various machine learning algorithms. This approach not only optimizes computational resources but also facilitates a deeper understanding of the complex interrelationships among water quality parameters. Through this pioneering research endeavor, the efficacy of ICA alongside PCA in addressing water quality index modeling challenges is evaluated. By integrating these techniques with machine learning methodologies, the study endeavors to provide actionable intelligence for agricultural stakeholders, aiding in informed decision-making and resource allocation. Moreover, by venturing into unexplored territory with the inclusion of ICA, the research contributes to expanding the methodological toolkit available for water quality assessment. As agriculture faces increasing pressure from climate change and resource scarcity, such innovative approaches hold promise in ensuring sustainable water management practices.

Keywords: Dimensionality reduction, Independent component analysis, Machine learning algorithms, Principal component analysis, Water quality index

Article Type: Research Article

Acknowledgment

We are grateful for the support of the University of Tabriz to conduct this research.

Conflicts of interest

The authors of this article declared no conflict of interest regarding the authorship or publication of this article.

Data availability statement

The datasets are available upon a reasonable request to the corresponding author.

Authors' contribution

Mohammad Taghi Sattari: Manuscript editing-original draft preparation, formal analysis and investigation; **Kimia Shirini:** Software, writing; **Sahar Javidan:** Formal analysis and investigation.

*Corresponding Author, E-mail: mtsattar@tabrizu.ac.ir

Citation: Sattari, M.T., Shirini, K., & Javidan, S. (2024). Evaluating the efficiency of dimensionality reduction methods in improving the accuracy of water quality index modeling using machine learning algorithms. *Water and Soil Management and Modelling*, *4*(2), 89-104.
DOI: 10.22098/mmws.2023.12434.1241

Received: 27 February 2023, Received in revised form: 02 April 2023, Accepted: 02 April 2023, Published online: 02 April 2023

Water and Soil Management and Modeling, Year 2024, Vol. 4, No. 2, pp. 89-104

Publisher: University of Mohaghegh Ardabili

مدل‌سازی و مدیریت آب و خاک

شاپا الکترونیکی: ۲۵۴۶-۲۷۸۳

ارزیابی کارائی روش‌های کاهش پارامترها در بهبود دقت مدل‌سازی شاخص کیفی آب در رودخانهٔ قزل اوزن با استفاده از الگوریتم‌های یادگیری ماشین

محمدتقی ستاری[۱*]، کیمیا شیرینی[۲]، سحر جاویدان[۳]

[۱] دانشیار، گروه علوم و مهندسی آب، دانشکدهٔ کشاورزی، دانشگاه تبریز، تبریز، ایران
[۲] دانشجوی دکتری، گروه مهندسی کامپیوتر، دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران
[۳] دانشجوی کارشناسی ارشد، گروه علوم و مهندسی آب، دانشکدهٔ کشاورزی، دانشگاه تبریز، تبریز، ایران

چکیده

آگاهی از کیفیت آب یکی از نیازهای مهم در برنامه‌ریزی، توسعه و حفاظت از منابع آب به‌شمار می‌رود. تعیین کیفیت آب برای مصارف مختلف از جمله آبیاری و شرب در مناطق مختلف ضروری است. استفاده از روش‌های مدرن داده‌کاوی، می‌توانند رویکرد مناسبی برای پیش‌بینی و طبقه‌بندی کیفیت آب ارائه دهند. در پژوهش حاضر کیفیت آب رودخانهٔ قزل اوزن در ایستگاه قره‌گونئی روستایی از توابع بخش حلب شهرستان ایجرود در استان زنجان مورد ارزیابی قرار گرفت. در این راستا شاخص کیفی آب شرب (WQI) با استفاده از پارامترهای شیمیایی سختی کل، قلیائیت (pH)، هدایت الکتریکی، کل مواد جامد محلول، کلسیم، سدیم، منیزیم، پتاسیم، کلر، کربنات، بی‌کربنات و سولفات در دورهٔ آماری ۲۱ساله (۱۳۹۸-۱۳۷۸) محاسبه شد. با توجه به تعداد نسبتاً زیاد پارامترها از روش‌های تحلیل مؤلفه‌های اصلی و تحلیل مؤلفه‌های مستقل برای کاهش ابعاد استفاده شد. سپس از الگوریتم‌های مختلف یادگیری ماشین شامل درخت تصمیم، رگرسیون لجستیک و شبکهٔ عصبی مصنوعی پرسپترون چندلایه برای مدل‌سازی شاخص کیفی آب استفاده شد. با استفاده از این روش‌ها تعداد پارامترهای مورد نیاز برای محاسبهٔ شاخص کیفی از ۱۲ به دو کاهش یافت. کاهش ابعاد داده‌ها باعث صرفه‌جویی در زمان نمونه‌برداری، پایش نمونه‌ها و تعیین کیفیت آب شده و هزینه‌های مورد نیاز برای مدل‌سازی را به مقدار قابل‌توجهی کاهش می‌دهد. نتایج نشان داد از بین روش‌های کاهش بعد روش تحلیل مؤلفه‌های اصلی نسبت به روش تحلیل مؤلفه‌های مستقل کارایی بهتری می‌تواند داشته باشد. هم‌چنین، نتایج نشان داد که از بین روش‌های مورد استفاده در مدل‌سازی، روش شبکهٔ عصبی پرسپترون چندلایه با استفاده از تحلیل مؤلفه‌های اصلی با ضریب تبیین ۰/۹۹، جذر میانگین مربعات خطا برابر ۴۴/۷۹ و ضریب ویلموت اصلاح شده برابر ۰/۹۹ بهترین عملکرد را داشته است. با توجه به این‌که ابعاد زیاد داده در بررسی و مدل‌سازی کیفیت آب باعث پیچیدگی و زمان بر بودن فرآیند مدل‌سازی می‌شود، لذا توصیه می‌شود از روش‌های کاهش بعد مانند تحلیل مولفه‌های اصلی برای کاهش ابعاد داده استفاده شود. نتایج حاصل از بررسی‌ها برتری روش تحلیل مؤلفه‌های اصلی نسبت به روش تحلیل مؤلفه‌های مستقل را نشان می‌دهد.

واژه‌های کلیدی: شاخص کیفی آب، کاهش ابعاد، الگوریتم‌های یادگیری ماشین، تحلیل مؤلفه‌های اصلی، تحلیل مؤلفه‌های مستقل

نوع مقاله: پژوهشی

[*] مسئول مکاتبات، پست الکترونیکی: mtsattar@tabrizu.ac.ir

استناد: ستاری، محمدتقی، شیرینی، کیمیا، و جاویدان، سحر (۱۴۰۳). ارزیابی کارائی روش‌های کاهش بعد در بهبود دقت مدل‌سازی شاخص کیفی آب با استفاده از الگوریتم های یادگیری ماشین. *مدل‌سازی و مدیریت آب و خاک*، ۴(۲)، ۸۹-۱۰۴.
DOI: 10.22098/mmws.2023.12434.1241

تاریخ دریافت: ۱۴۰۱/۱۲/۰۸، تاریخ بازنگری: ۱۴۰۲/۰۱/۱۳، تاریخ پذیرش: ۱۴۰۲/۰۱/۱۳، تاریخ انتشار: ۱۴۰۲/۰۱/۱۳
مدل‌سازی و مدیریت آب و خاک، سال ۱۴۰۳، دوره ۴، شماره ۲، صفحه ۸۹ تا ۱۰۴
ناشر: دانشگاه محقق اردبیلی

۱- مقدمه

منابع آب، اعم از آب‌های سطحی و زیرزمینی به‌طور گسترده مورد بهره‌برداری قرار گرفته‌اند و از این‌رو در حال حاضر با مشکلات جدی آلودگی و کمبود در سراسر جهان مواجه هستند. بنابراین، توجه جدی به بهبود و حفظ کیفیت و کمیت آن‌ها ضروری است. از این‌رو نیاز به توسعهٔ روش‌های مؤثر برای ارزیابی منابع آب‌های زیرزمینی و سطحی برای توسعهٔ پایدار و ایمنی سلامت انسان مطرح شده است (Boyacioglu, 2007). آلودگی هر دو شکل منابع آبی به دلایل متعددی مانند رواناب‌های کشاورزی، آلودگی خانگی و صنعتی بسیار رایج است (Bailey and Solomon, 2004). به‌طور سنتی، آب‌های سطحی آسان‌ترین منبع برای مصارف عمومی و حساسیت زیادی در برابر آلودگی‌ها دارند. بنابراین، از آن جایی‌که بیماری‌های قابل انتقال از طریق آب به‌عنوان ۱۰ عامل اصلی شناخته شده در مرگ و میر در سراسر جهان را تشکیل می‌دهند یک رویکرد دقیق و آگاهانه برای نظارت و ارزیابی سلامت آب‌های سطحی برای اطمینان از سلامت آن مورد نیاز است (Massoud et al., 2012). پارامترهای شیمیایی، فیزیکی و بیولوژیکی متنوعی برای کیفیت آب وجود دارد. پژوهش‌گران متعددی شاخص کیفیت آب را در قالب یک عبارت ساده برای نمایش کیفیت عمومی آب‌های سطحی پیشنهاد کرده‌اند که روشی مختصر و جامع برای بیان کیفیت آب برای مراحل مختلف مصرف است (Zeinalzadeh et al., 2017). این عدد نشان‌دهندهٔ کیفیت یا وضعیت آلودگی آب با تجمیع مقادیر پارامترهای مختلف است (Gorde and Jadhav, 2013).

امروزه پژوهش‌های مربوط به کیفیت آب به‌دلیل افزایش آب‌های سطحی سرعت بیش‌تری گرفته است (Icaga, 2007). چندین شاخص کیفیت آب[1] (WQI) در سطح جهانی به‌منظور نظارت بر کیفیت آب شیرین برای مصرف مستقیم انسان و سایر مصارف توسعه یافته است (Jie et al., 2016). کیفیت آب در اکوسیستم‌های آبی توسط چندین پارامتر بیولوژیکی، فیزیکی و شیمیایی تعیین می‌شود. کیفیت آب تغییرات زیادی را براساس مکان و زمان نشان می‌دهد. بنابراین، پایش منظم آن منجر به تولید یک ماتریس دادهٔ پیچیده و بزرگ متشکل از تعداد زیادی پارامتر می‌شود که درک آن‌ها به‌علت افزایش پیچیدگی مسأله عمدتاً دشوار است. علاوه‌براین همواره دسترسی به تمامی این پارامترها ممکن نیست. استفاده از تکنیک‌های مختلف آماری چندپارامتره، مانند تحلیل مؤلفه‌های اصلی[2] (PCA) و تحلیل مؤلفه‌های مستقل[3] (ICA) به تفسیر بهتر نتیجه کمک می‌کند و باعث حل مسأله با تعداد کم پارامترهای مؤثر می‌شود (Jie et al., 2016).

در پژوهشی، Dezfooli et al. (2017) به طبقه‌بندی کیفی آب رودخانهٔ کارون بر اساس حداقل پارامترهای کیفی پرداختند. نتایج مطالعهٔ آن‌ها نشان داد که روش شبکهٔ عصبی احتمالی با استفاده از پارامترهای کیفی کدورت[4]، کلیفرم مدفوعی[5] و کل مواد جامد با دقت ۹۰/۷۸ می‌تواند به طبقه‌بندی کیفی آب بپردازد. در پژوهش دیگری، Khalili et al. (2021) کیفیت آب رودخانهٔ تالار استان مازندران را با استفاده از ترکیب شاخص‌های کیفیت آب ارزیابی کرده و به مدل‌سازی چندپارامتره پرداختند. هدف پژوهش آن‌ها ارزیابی کیفیت آب و بررسی مکانیسم‌های کنترل‌کنندهٔ آن با استفاده از شاخص کیفیت آشامیدنی و تعیین تیپ شیمیایی آب با استفاده از نمودار سه خطی پایپر بود. هم‌چنین، Soleimanpour et al. (2018) برای تعیین مؤثرترین عامل کیفیت آب آشامیدنی دشت کازرون از تکنیک طبقه‌بندی و رگرسیون درختی استفاده کردند. نتایج نشان داد که دو پارامتر کل جامدات محلول و مقدار کلسیم بر کیفیت آب آشامیدنی، تأثیر بیش‌تری داشته است که علت آن را ساختار سازندهای زمین‌شناسی منطقه و وجود کربنات کلسیم در ترکیب آن‌ها بیان کرده‌اند. در ادامه، Al-Mukhtar and Al-Yaseen (2019) توانایی سه مدل مختلف از تکنیک‌های هوش مصنوعی از جمله سیستم استنتاج فازی مبتنی بر عصبی تطبیقی[6] (ANFIS)، شبکه‌های عصبی مصنوعی (ANN) و مدل رگرسیون چندگانه (MLR) را برای پیش‌بینی و تخمین TDS و EC مورد بررسی قرار دادند. نتایج پژوهش ایشان نشان داد که ANFIS نسبت به سایر روش‌های ارزیابی، کارایی بالایی داشته و بهترین تناسب را با داده‌های مشاهداتی نشان

[1] Water quality index

[2] Principal component analysis
[3] Independent components analysis
[4] Turbidity
[5] Fecal coliform
[6] Adaptive neural based fuzzy inference system

داد. در رودخانهٔ گانگا هند نیز Tripathi and Singal (2019) برای انتخاب پارامتر مؤثر برای توسعهٔ یک شاخص کیفیت آب جدید از روش تحلیل مؤلفهٔ اصلی استفاده کردند و تعداد پارامترهای مورد نیاز برای محاسبهٔ شاخص کیفی را از ۲۸ به نُه کاهش دادند. آن‌ها بیان کردند که این‌کار باعث صرفه‌جویی زمان، تلاش و هزینهٔ موردنیاز برای نظارت نمونه‌ها در تعداد زیادی از پارامترها می‌شود.

از سایر پژوهش‌ها، Chen et al. (2020) به تحلیل مقایسه‌ای عملکرد پیش‌بینی کیفیت آب سطحی در چین و شناسایی پارامترهای کلیدی آب با استفاده از مدل‌های مختلف یادگیری ماشین براساس داده‌های بزرگ پرداختند. آن‌ها برای پایش کیفیت آب در آینده و ارائهٔ هشدار به موقع کیفیت آب، روش درخت تصمیم، جنگل تصادفی و جنگل آبشار عمیق[1] را در اولویت قرار دادند. در رودخانهٔ آلاداغ کشور ترکیه نیز Sattari et al. (2021) طبقه‌بندی کیفیت آب سطحی برای مصارف آبیاری و کشاورزی با استفاده از روش‌های داده‌کاوی را بررسی کردند. آن‌ها نتیجه گرفتند که در طبقه‌بندی مبتنی بر USSL[2]، روش‌های داده‌کاوی دارای یک خطا، سیستم‌های شوفیلد[3] ۱۹۳۳ و ۱۹۳۵ دارای شش خطا و طبقه‌بندی بر اساس روش ویلکاکس بدون خطا بوده است. برای طبقه‌بندی و پیش‌بینی Islam Khan et al. (2021) پژوهشی را بر اساس رگرسیون مؤلفه اصلی و رویکرد طبقه‌بندی‌کنندهٔ تقویت گرادیان، کیفیت آب انجام دادند. آن‌ها رویکرد طبقه‌بندی‌کنندهٔ تقویت گرادیان را با دقت طبقه‌بندی صد در صد، به‌عنوان روش برتر معرفی کردند و دریافتند که این روش در مقایسه با مدل‌های پیشرفته، عملکرد مطلوبی داشته است. برای ارزیابی کیفت آب در جنوب غربی نیجریه نیز Johnson et al. (2021) روش منطق فازی را با روش‌های شاخص کیفیت آب مقایسه کرده‌اند. آن‌ها BOD5 را به‌عنوان پارامتری با تأثیر بیش‌تر برای ارزیابی کیفیت آب سطحی معرفی کرده‌اند. هم‌چنین، Khoi et al. (2022) برای پیش‌بینی شاخص کیفیت آب در رودخانهٔ لابوونگ ویتنام، از مدل‌های یادگیری ماشین استفاده کردند. نتایج پژوهش آن‌ها نشان داد که مدل تقویت گرادیان شدید (XGBoost) با R=۰/۹۸۹ و RMSE[4]=۰/۱۰۷، عملکرد خوبی در پیش‌بینی WQI داشته است. در نهایت، Yusri et al. (2022) به طبقه‌بندی کیفیت آب با استفاده از روش ماشین بردار پشتیبان و افزایش گرادیان شدید پرداختند. آن‌ها یک الگوریتم طبقه‌بندی برای پیش‌بینی طبقه‌بندی کیفیت آب ایجاد کردند و نتیجه گرفتند که مدل XGBoost با دقت ۹۴ درصد در مقایسه با مدل SVM با دقت ۶۷ درصد، کارایی بهتری داشته است.

با توجه به این‌که رودخانهٔ قزل اوزن از رودخانه‌های مهم استان زنجان و یکی از بلندترین رودخانه‌های ایران است و توزیع آب چندین شهر از این رودخانه تأمین می‌شود بنابراین، اهمیت تشخیص شاخص کیفی آب این رود بسیار ضروری است. لذا، هدف مقالهٔ حاضر بررسی کیفیت شیمیایی آب رودخانهٔ قزل اوزن در ایستگاه قره‌گونئی روستایی از توابع بخش حلب شهرستان ایجرود در استان زنجان در دورهٔ آماری ۲۱ساله (۱۳۹۸-۱۳۷۸) بر اساس شاخص کیفی آب با استفاده از داده‌های مربوط به پارامترهای کیفی آب است. به‌دلیل عدم ثبت و عدم دسترسی به داده‌های بلند مدت، از داده‌های ۲۱ سال استفاده شد. قطعا طولانی بودن طول دورهٔ آماری به نتایج بهتر و دقیق‌تر منجر خواهد شد. به‌منظور کاهش پارامترهای ورودی مدل‌های یادگیری ماشینی از دو تکنیک آماری خطی تحلیل مؤلفه‌های اصلی (PCA) و تحلیل مؤلفهٔ مستقل (ICA) استفاده شده و عملکرد آن‌ها مورد ارزیابی قرار گرفت. سپس از الگوریتم‌های یادگیری ماشین شامل درخت تصمیم، رگرسیون لجستیک و شبکهٔ عصبی پرسپترون چندلایه برای مدل‌سازی شاخص کیفی آب استفاده شده است.

۲- مواد و روش‌ها

۲-۱- منطقهٔ مورد مطالعه

استان زنجان در بخش شمال غربی کشور ایران واقع شده و مساحت آن حدود ۱۲۶۴۳۲۴ کیلومتر مربع است. قره‌گونئی روستایی از توابع بخش حلب شهرستان ایجرود در استان زنجان است. ایستگاه آب‌سنجی قره‌گونئی روی رودخانهٔ قزل اوزن در عرض‌های جغرافیایی۳۴ درجه و ۵۵ دقیقه تا ۳۷ درجه و ۵۵ دقیقهٔ عرض شمالی و ۴۶ درجه و ۲۷ دقیقه تا ۴۹ درجه و ۲۰ دقیقهٔ طول شرقی

[1] Deep cascade forest
[2] United states salinity laboratory
[3] Schofield
[4] Root mean square error

گسترده شده و در ارتفاع ۱۴۲۸ متری از سطح دریا واقع شده است. رودخانهٔ قزل اوزن از ارتفاعات چهل‌چشمهٔ استان کردستان سرچشمه گرفته و طول آن به بیش از ۵۵۰ کیلومتر می‌رسد این رودخانه از استان‌های زنجان، آذربایجان‌شرقی و اردبیل عبور می‌کند. پس از جذب رودخانه‌های متعدد در جنوب استان گیلان به رودخانهٔ شاهرود پیوسته و سفیدرود را تشکیل می‌دهد و در نهایت به دریای خزر می‌ریزد. این رودخانه به‌دلیل ختم شدن قزل اوزن به دریاچهٔ سد شهریار در انتهای حوضهٔ آذربایجان‌شرقی و تأثیر آن بر دریاچهٔ سفیدرود و زیست‌بوم این دو آبگیر، حائز اهمیت بالایی است. از دیدگاه ژئومورفولوژی، رودخانه‌های آبرفتی با بستری سنگلاخی، رسی–سیلتی و ماسه‌ای است و تحت تأثیر شرایط خاص هیدرولیکی، توپوگرافی، مواد سازندهٔ بستر و کناره‌ها قرار دارد. این رودخانه از رودخانه‌های مهم استان زنجان و یکی از بلندترین رودخانه‌ها ی ایران می‌باشد. تأمین آب شرب و کشاورزی چندین شهر از طریق این رودخانه انجام می‌شود. رودخانهٔ قزل اوزن در مسیر خود از کاربری‌های مختلفی چون اراضی کشاورزی، مسکونی، و صنعتی می‌گذرد. موقعیت جغرافیای منطقهٔ مورد مطالعه و ایستگاه قره گونئی در شکل ۱ ارائه شده است.

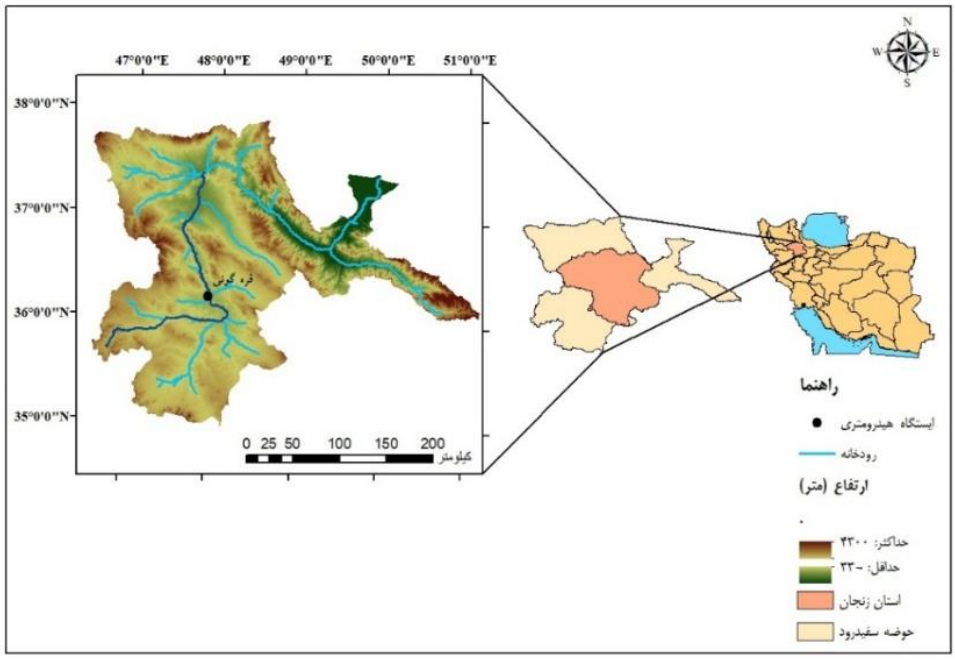
برای محاسبهٔ شاخص WQI داده‌های پارامترهای شیمیایی سختی کل (TH)، هدایت الکتریکی (EC)، قلیائیت (pH)، کربنات (CO_3)، بی‌کربنات (HCO_3)، کلر (Cl)، سولفات (SO_4)، کلسیم (Ca)، منیزیم (Mg)، سدیم (Na)، پتاسیم (K) و کل مواد جامد محلول (TDS) طی سال‌های ۱۳۷۸ تا ۱۳۹۸ استفاده شد. در جدول ۱ مشخصات آماری پارامترهای مورد استفاده در ایستگاه قره‌گونئی ارائه شده است.

شاخص WQI یکی از شاخص‌های پرکاربرد برای تعیین کیفیت شیمیایی آب‌های سطحی و زیرزمینی است که در سه مرحله محاسبه و استفاده می‌شود. در مرحلهٔ اول، وزن (W_i) هر پارامتر کیفیت آب به‌دلیل اهمیت آن برای آب آشامیدنی اندازه‌گیری و وزن نسبی (W_i) با استفاده از رابطهٔ (۱) به‌دست می‌آید که در آن n تعداد پارامترهاست.

$$W_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \qquad (۱)$$

در مرحلهٔ دوم، رتبه‌بندی کیفیت (q_i) هر پارامتر مشخص و نسبت مقدار استاندارد آن بر اساس قوانین سازمان جهانی بهداشت اندازه‌گیری می‌شود (رابطهٔ ۲). در این رابطه، C_i غلظت پارامترهای شیمیایی نمونه‌های آب و S_i استاندارد آب آشامیدنی سازمان بهداشت جهانی هر پارامتر (میلی‌گرم در لیتر) هستند.

$$q_i = \left(\frac{c_i}{S_i}\right) \times 100 \qquad (۲)$$



شکل ۱- موقعیت مکانی منطقهٔ مورد مطالعه

Figure 1- Location of the studied area

جدول ۱- مشخصات آماری پارامترهای مورد مطالعه در ایستگاه قره‌گونئی طی سال‌های ۱۳۷۸ تا ۱۳۹۸

Table 1 - Statistical characteristics of the studied parameters in Qara Gunei station during the years 2000 to 2020

Statistic	Unit	Min	Max	Mean	Variance	Standard deviation	Skewness	Kurtosis	coefficient of variation (CV)
TDS	میلی‌گرم بر لیتر	153.90	125370.00	7721.4	147988331.64	12165.05	6.34	56.92	1.58
EC	میکروموس بر سانتی‌متر	660.00	19900.00	12532.01	37174249.76	19280.63	6.32	56.89	1.54
pH		7.00	8.23	7.78	0.05	0.22	-0.59	0.54	0.03
CO_3		0.0	0.11	0	0.00	0.01	11.50	134.13	8.98
HCO_3		1.80	16.00	4.10	2.51	1.58	4.08	24.88	0.39
Cl		1.85	480.90	100.73	13142.18	114.64	1.35	0.86	1.14
SO_4	میلی‌اکی‌والان در لیتر	0.0	49.96	6.78	51.13	7.15	3.33	13.52	1.05
Ca		0.88	1876.00	21.65	23467.56	153.19	12.02	143.05	7.07
Mg		1.45	67.31	10.24	106.99	10.34	2.91	11.34	1.01
Na		2.04	438.90	90.04	10305.99	101.52	1.44	1.38	1.13
K		0.0	1.96	0.41	0.14	0.37	1.46	1.73	0.90
TH	میلی‌گرم بر لیتر	187.50	5805.50	967.03	808169.10	898.89	2.44	8.52	0.93

در مرحلهٔ سوم، WQI به‌عنوان شاخص کیفی آب شرب اندازه‌گیری می‌شود (رابطهٔ ۳) و به پنج کلاس عالی، خوب، ضعیف، بسیارضعیف و نامناسب برای آشامیدن تقسیم‌بندی می‌شوند (جدول ۲).

$$WQI = \sum_{i=1}^{n} W_i\, q_i \qquad (۳)$$

جدول ۲- طبقه‌بندی کیفیت آب بر اساس ارزش WQI

Table 2- Classification of water quality based on WQI value

نوع آب	کلاس	دامنهٔ WQI
آب با کیفیت عالی	I	< 50
آب با کیفیت خوب	II	50-100
آب با کیفیت ضعیف	III	100-200
آب با کیفیت خیلی ضعیف	IV	200-300
غیرقابل آشامیدن	V	> 300

برای محاسبهٔ WQI از ۱۲ پارامتر شیمیایی شامل کل جامدات محلول، هدایت الکتریکی، سولفات، سختی کل، pH، کلر، سدیم، پتاسیم، منیزیوم، کلسیم، کربنات و بی‌کربنات استفاده می‌شود. استانداردهای آب آشامیدنی برای هر پارامتر شیمیایی مطابق با دستورالعمل WHO در جدول ۳ ذکر شده است. وزن هر پارامتر با توجه به اهمیت نسبی آن در کیفیت آب برای اهداف نوشیدن مشخص می‌شود. حداکثر وزن پنج به مجموع جامدات محلول (TDS) و EC مربوط می‌شود. وزن چهار به SO_4 و TH اختصاص داده شده است. وزن سه برای pH، کلر و Na و وزن دو برای K، Mg، Ca، CO_3 و HCO_3 در نظر گرفته می‌شود. با توجه به جدول ۲ می‌توان نتیجه گرفت که دامنهٔ تغییرات اعداد در پارامترهای شیمیایی با هم متفاوت است و تفاوت زیادی با هم دارند. بنابراین، هنگام استفاده از این داده‌ها برای انجام مدل‌سازی، نسبت به نرمال کردن آن‌ها با استفاده از رابطهٔ (۴) اقدام شد.

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (۴)$$

در رابطهٔ (۴)، X_n داده‌های نرمال شده، X داده‌های واقعی، X_{min} کم‌ترین مقدار دادهٔ مورد بررسی و X_{max} بیش‌ترین مقدار دادهٔ مورد مطالعه است. تعداد داده‌های مورد بررسی در پژوهش حاضر نیز ۱۴۹ داده بود.

جدول ۳- بازهٔ مطلوب هر پارامتر کیفیت به همراه وزن نسبی آن (World Health Organization, 2010)

Table 3- The optimal range of each quality parameter along with its relative weight (World Health Organization, 2010)

پارامتر	حد مطلوب (WHO) (میلی‌گرم بر لیتر)	حد مجاز (WHO) (میلی‌گرم بر لیتر)	وزن (Wi)	وزن نسبی (W_i)
TDS	500	1000	5	0.135
pH	6.5-8.5	8.5	3	0.081
EC	1500	1500	5	0.135
TH	300	600	4	0.108
Ca	75	75	2	0.054
Na	200	200	3	0.081
Mg	30	30	2	0.054
K	10	10	2	0.054
CL	200	200	3	0.081
CO3	100	100	2	0.054
SO4	200	200	4	0.081
HCO3	100	100	2	0.054
	مجموع		37	1.000

۲-۲- روش‌های کاهش بعد

کاهش بعد به این معنی است که یک داده را از فضایی با ابعاد بالا به یک فضای جدید با ابعاد پایین‌تر منتقل کنیم به‌طوری‌که عمده‌ترین اطلاعات داده باقی بماند. در این مقاله از دو روش تحلیل مؤلفهٔ اساسی و روش تحلیل مؤلفه‌های مستقل استفاده شده است. برای پیاده‌سازی از نرم‌افزار پایتون استفاده شد.

۲-۲-۱- تحلیل مؤلفه‌های اساسی

یکی از کاربردهای اصلی روش تحلیل مؤلفه‌های اصلی کاهش ویژگی است. این الگوریتم می‌تواند مؤلفه‌های اصلی را شناسایی و تنها ویژگی‌هایی را که ارزش بیش‌تری دارند، مشخص کند. بردار X را در نظر بگیرید. می‌توان فرض کرد که مؤلفه‌های آن دارای میانگین صفر هستند و اگر نباشد به‌راحتی میانگین را حساب و از آن‌ها کم می‌شود (Daffertshofer et al., 2004). بردارهای ویژگی جدید به کمک رابطهٔ (۵) به‌دست می‌آید.

$$C = P.x \qquad (۵)$$

جدول ٤- پارامترهای مورد استفاده در PCA و ICA

Table 4- Variables used in PCA and ICA

ماتریس کواریانس X	Y
بردار مورد نظر	X
ماتریس بردار ویژه	A
ماتریس مقادیر ویژه از ماتریس کواریانس Y	B
هر عضو A	f
هر عضو B	l
معکوس A	W

۲-۲-۲- تحلیل مؤلفه‌های مستقل

تحلیل مؤلفه‌های مستقل یک تکنیک آماری و محاسباتی برای آشکار کردن عوامل پنهانی است که زمینه‌ساز مجموعه‌ای از پارامترهای تصادفی، اندازه‌گیری‌ها یا امواج هستند. الگوریتم تحلیل مؤلفه‌های مستقل ICA در حل مسأله جداسازی موج‌های منبع یکی از مهم‌ترین مسائلی است که در دههٔ اخیر توجه زیادی از پژوهش‌گران شاخه‌های مختلف علوم مهندسی را به خود جلب کرده است. در این تحلیل خواص آماری مراتب بالا برای جداسازی امواج به‌کار گرفته می‌شود. در چند سال اخیر علاقهٔ زیادی به استفاده از روش‌های هندسی در ICA برای به‌دست‌آوردن یک روش بهینه‌سازی و بهبود آن نشان داده شده است. فرض کنیم که ترکیب خطی از مؤلفه مستقل را داریم و $x1, x2, \ldots, xn$ متغیرهای تصادفی باشند. در نتیجه $x(t)$ یک نمونه از این متغیر تصادفی است. بنابراین، عبارت بالا را می‌توان به‌صورت رابطهٔ (۶) نشان داد. اگر ستون ماتریس A باشد، a_j به‌عنوان یک ستون ماتریس A خواهد بود (رابطهٔ ۷).

$$x = As \qquad (۶)$$

$$x = \sum_{i=1}^{n} a_i s_i \qquad (۷)$$

مدل آماری رابطهٔ (۶) مدل تحلیل مؤلفه‌های مستقل (ICA) خوانده می‌شود که بیان‌کنندهٔ این است که چگونه داده‌های به‌دست آمده از روی مؤلفه‌های مستقل به‌صورت ترکیبی ساخته می‌شوند. مؤلفه‌های مستقل به‌عنوان متغیرهای پنهان هستند؛ یعنی این‌که به‌طور مستقیم دیده نمی‌شوند. هم‌چنین، ماتریس ترکیب معلوم نیست و تنها چیزی که ما در دست داریم مقادیر مشاهده x است که مقادیر A و s از روی x تخمین زده می‌شوند. نقطهٔ شروع برای تخمین A و s فرض سادهٔ i است که مؤلفه‌های S از لحاظ آماری مستقل هستند. هنگامی‌که ماتریس A تخمین زده شد، معکوس آن محاسبه می‌شود و با استفاده از آن مؤلفه‌های مستقل به‌صورت رابطهٔ (۸) محاسبه می‌شود.

$$s = Wx \qquad (۸)$$

حال می‌توان ماتریس W را به‌گونه‌ای انتخاب کرد تا کاهش بعد نیز ارائه گشته است. در این پژوهش برای اولین‌بار از این روش برای کاهش ابعاد ورودی مدل‌های یادگیری ماشین در پیش‌بینی شاخص کیفی آب استفاده می‌شود. روش تحلیل مولفه‌های مستقل (ICA) روش جدیدی که بیش‌تر در پردازش امواج استفاده می‌شده است، می‌کوشد تا خطای بار داده‌ها را کم کرده و نتایج با بهترین دقت را ارائه دهد.

۲-۳- مدل‌های یادگیری ماشین

۲-۳-۱- درخت تصمیم

درخت تصادفی یک طبقه‌بندی‌کنندهٔ تحت نظارت است و از یک ایدهٔ جمع‌آوری برای تولید مجموعه‌ای تصادفی از داده‌ها برای ساخت درخت تصمیم استفاده می‌کند. درختان تصادفی اساساً ترکیبی از دو الگوریتم موجود در یادگیری ماشین هستند که از

درختان تک مدل با ایده‌های جنگل تصادفی ترکیب شده‌اند. درختان نمونه، درختان تصمیم‌گیری هستند که در آن هر برگ یک مدل خطی دارد که برای زیرفضای محلی توصیف شده توسط این برگ بهینه شده است. جنگل‌های تصادفی نشان داده‌اند که عملکرد درختان تصمیم واحد را به میزان قابل توجهی بهبود می‌بخشند (Kalmegh, 2015). درختان ناپایدار هستند؛ یعنی تغییرات کوچک در داده‌های آموزشی می‌تواند منجر به ساخت درختانی شود که از نظر ساختار بسیار متفاوت هستند. اگرچه این ممکن است برای یک درخت واحد مشکل‌ساز باشد، اما می‌توان از این اثر در یک گروه استفاده کرد. الگوریتم درخت تصادفی می‌تواند با مشکلات طبقه‌بندی و رگرسیون مقابله کند. طبقه‌بندی‌کنندهٔ درختان تصادفی، بردار ویژگی ورودی را شامل می‌شوند. آن را با هر درخت در جنگل طبقه‌بندی می‌کند و برچسب کلاس را که اکثریت رأی را دریافت کرده است، نشان می‌دهد. در مورد رگرسیون، پاسخ طبقه‌بندی‌کننده، میانگین پاسخ‌ها به تمام درختان جنگل است. همهٔ درختان با پارامترهای یکسان اما در مجموعه‌های آموزشی مختلف آموزش می‌بینند (Ajayram et al., 2021). برای ساختن یک درخت تصادفی، سه انتخاب اصلی وجود دارد که عبارتند از: روش تقسیم برگ‌ها، نوع پیش‌بینی‌کننده برای استفاده در هر برگ و روش تزریق تصادفی به درختان. یک تکنیک رایجی که برای معرفی تصادفی‌بودن در یک درخت می‌توان به آن اشاره کرد. ساخت هر درخت با استفاده از یک مجموعه داده بوت استرپ یا زیر نمونه‌برداری است. به این ترتیب، هر درخت در جنگل، روی داده‌های کمی متفاوتی آموزش داده می‌شود که تفاوت‌های بین درختان را معرفی می‌کند (Denil et al., 2014).

۲-۳-۲- رگرسیون لجستیک

رگرسیون لجستیک[1] یکی از الگوریتم‌های یادگیری ماشین است. این الگوریتم برای مسائل طبقه‌بندی استفاده می‌شود که در آن پارامتر وابستهٔ گسسته[2] مطرح می‌شود (La Valley et al., 1999). رگرسیون لجستیک یک مدل آماری رگرسیون برای پارامترهای وابسته دوسویی مانند بیماری یا سلامت، مرگ یا زندگی است. این مدل را می‌توان به‌عنوان مدل خطی تعمیم‌یافته‌ای که از تابع لوجیت به‌عنوان تابع پیوند استفاده می‌کند و خطای آن از توزیع چندجمله‌ای پیروی می‌کند، به‌حساب‌آورد. رگرسیون لجستیک می‌تواند یک مورد خاص از مدل خطی عمومی و رگرسیون خطی دیده شود. مدل رگرسیون لجستیک، بر اساس فرض‌های کاملاً متفاوتی (رابطهٔ پارامترهای وابسته و مستقل) از رگرسیون خطی است. تفاوت مهم این دو مدل در دو ویژگی رگرسیون لجستیک می‌تواند دیده شود.

۲-۳-۲- شبکهٔ عصبی مصنوعی پرسپترون چندلایه

برای پیش‌بینی کیفیت آب از الگوریتم پرسپترون چندلایه که دسته‌ای از شبکه‌های عصبی مصنوعی پیشخور است استفاده شد. یک MLP[3] شامل حداقل سه‌لایهٔ گره است: لایهٔ ورودی، لایهٔ پنهان و لایهٔ خروجی. به‌جز گره‌های ورودی، هر گره یک نورون است که از یک تابع فعال‌سازی غیرخطی استفاده می‌کند. پرسپترون چندلایه یک کلاس کاملاً متصل از شبکهٔ عصبی مصنوعی پیش‌خور[4] (ANN) است. اصطلاح MLP به‌طور مبهم استفاده می‌شود، گاهی‌اوقات به‌طور آزاد به معنای هر ANN پیش‌رونده، گاهی‌اوقات به‌طور دقیق به شبکه‌های متشکل از چندین لایهٔ پرسپترون (با فعال‌سازی آستانه) اشاره می‌کند (Pinkus, 1999).

MLPs برای مسائل پیش‌بینی طبقه‌بندی مناسب هستند که در آن ورودی‌ها یک کلاس یا برچسب تخصیص داده می‌شوند. آن‌ها برای مسائل پیش‌بینی رگرسیون مناسب هستند که در آن یک کمیت باارزش واقعی با توجه به مجموعه‌ای از ورودی‌ها پیش‌بینی می‌شود. پرسپترون چندلایه یکی از رایج‌ترین مدل‌های شبکهٔ عصبی مورد استفاده در زمینهٔ یادگیری عمیق است. MLP ساده‌تر از مدل‌های پیچیدهٔ امروزی است. با این‌حال، تکنیک‌های معرفی شده راه را برای شبکه‌های عصبی پیشرفته‌تر هموار می‌کند.

[1] Logistic regression
[2] Categorical
[3] Multilayer perceptron
[4] Feedforward neural network

۲-٤- معیارهای ارزیابی

۲-٤-۱- نمودار تیلور

باتوجه به رقابت شدید مدل‌های یادگیری ماشین در مدل‌سازی شاخص کیفی آب جهت انتخاب مدل برتر از معیارهای عددی و گرافیکی استفاده می‌شود. نمودار تیلور یکی از معیارهای گرافیکی شناخته شده جهت سنجش دقت مدل‌های پیش‌بینی است. در این پژوهش با توجه به رقابت مدل‌های پیش‌بینی مورد استفاده جهت درک صحیح و رسیدن به شناخت کافی در انتخاب مدل برتر از نمودار تیلور نیز استفاده شد. در نمودار تیلور، ضریب همبستگی و اختلاف ریشهٔ میانگین مربع بین دو فیلد، به همراه نسبت انحراف معیار دو الگو، همه با یک نقطه در یک فضای دوبعدی نشان داده می‌شود. تمامی نمودارها به وسیلهٔ نرم‌افزار پایتون پیاده‌سازی شده است. این آمارها با هم خلاصه‌ای سریع از درجهٔ مطابقت الگو را ارائه می‌دهند و می‌توان میزان دقت یک مدل شبیه‌سازی سیستم طبیعی را اندازه‌گیری کرد. این نمودار به‌ویژه در ارزیابی شایستگی‌های نسبی مدل‌های رقیب و در نظارت بر عملکرد کلی به‌عنوان یک مدل مفید در نظر گرفته می‌شود. بنابراین، برای داشتن تصور بصری بهتر از نتایج به‌دست آمده از مدل‌سازی می‌توان نمودار تیلور را رسم کرد. در حالت کلی، نمودارها به‌ویژه در ارزیابی جنبه‌های مختلف مدل‌های پیچیده یا در سنجش مهارت نسبی بسیاری از مدل‌های مختلف مفید هستند (Taylor, 2001).

۲-٤-۲- نمودار ویولن [1]

نمودار ویولن برای بصری‌سازی توزیع داده‌های عددی و تراکم احتمالی آن‌ها به‌منظور مقایسهٔ داده‌های آماری به‌صورت خلاصه (مانند بازه‌ها و چارک‌ها) استفاده می‌شود و تغییرات و اختلافات داده‌ها را نشان می‌دهد (Hintze and Nelson, 1998).

۲-٤-۳- شاخص ضریب همبستگی

ضریب همبستگی معیاری آماری از قدرت رابطهٔ خطی بین دو پارامتر و یا مقادیر مشاهداتی و محاسبه شده از مدل است. مقادیر آن می‌تواند از منفی یک تا یک پارامتر باشد. ضریب همبستگی منفی یک همبستگی منفی یا معکوس کامل را توصیف می‌کند. در رابطهٔ (۹) y_i مقدار برآورد شده از مدل، x_i مقدار محاسبه شده از شاخص کیفی آب و N تعداد داده‌ها هستند.

$$R = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 . \sum_{i=1}^{N}(y_i - \overline{y})^2}} \quad (۹)$$

۲-٤-٤- جذر میانگین مربعات خطا

جذر میانگین مربعات خطا (RMSE) یکی از معیارهای اندازه‌گیری خطا است که برای ارزیابی مدل‌های مختلف استفاده می‌شود (Coutsias et al., 2004). برای محاسبهٔ RMSE می‌توان از رابطهٔ (۱۰) استفاده کرد.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \quad (۱۰)$$

در رابطهٔ (۱۰)، y_i مقدار برآورد شده از مدل، x_i مقدار محاسبه شده از شاخص کیفی آب و N تعداد داده‌ها هستند.

۲-٤-۵- ضریب ویلموت اصلاح شده

معادلهٔ اصلی ضریب ویلموت به‌صورت رابطهٔ (۱۱) است. خطاهای بزرگ‌تر، زمانی‌که مجذور می‌شوند، تأثیر آن خطاها را بر مجموع خطاهای مربعی بیش‌تر می‌کنند. با این وجود، Willmott at al. (2012) نسخه‌ای از d را ارائه کردند که بر اساس مجموع مقادیر مطلق خطاها بود و آن را d_1 نامیده‌اند (Willmott at al., 2012).

$$WI = \left|1 - \left[\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(|y_i - \overline{x}| + |x_i - \overline{x}|)^2}\right]\right|, 0 \leq WI \leq 1 \quad (۱۱)$$

$$\lambda_1 > \lambda_2 > \cdots > \lambda_m > \cdots > \lambda_n \quad (۱۲)$$
$$P = [\emptyset_1, \emptyset_2, \ldots, \emptyset_m]$$

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n \quad (۱۳)$$

شمای کلی مراحل انجام پژوهش در شکل ۲ نمایش داده شده است.

[1] Violin plot

شکل ۲- نمودار مراحل انجام مدل‌سازی شاخص WQI

Figure 2- Flowchart of WQI index modeling steps

۳- نتایج و بحث

مقادیر مشخصات آماری بخش صحت‌سنجی برای داده‌های مشاهداتی و روش‌های یادگیری ماشین در جدول ۵ آورده شده است. به‌طور کلی با توجه به نتایج جدول‌های ۵ و ل۶ و شکل ۳ نتیجه گرفته شد که الگوریتم پرسپترون چندلایه بیش‌ترین تأثیر را در مدل‌سازی شاخص کیفی آب داشته است که به‌عنوان معیار ارزیابی مورد بررسی قرار گرفته‌اند.

نتایج نشان می‌دهد روش PCA یکی از روش‌های بسیار مناسب نسبت به روش ICA برای کاهش بعد ویژگی‌ها است. هم‌چنین، با مقایسهٔ جدول ۵ می‌توان نتیجه گرفت برای پیش‌بینی کیفیت آب روش شبکهٔ عصبی مصنوعی چندلایه دقت بالا و خطای کمی نسبت به روش‌های اشاره شده توسط سایر روش‌های یادگیری ماشین دارد. روش LR از بین روش‌های اشاره شده در یادگیری ماشین دقت بسیار بهتری دارد. بنابراین، می‌توان نتیجه گرفت روش‌های یادگیری ماشین نسبت به روش پرسپترون چندلایه دقت کم‌تری دارد. علاوه‌بر این در بین روش‌های اشاره شده توسط یادگیری ماشین روش‌های LR و DTree به‌ترتیب بیش‌ترین دقت و کم‌ترین خطا را در پیش‌بینی شاخص کیفیت آب دارند.

جدول ۵- مشخصات آماری مشاهداتی و تخمینی از مدل‌ها برای بخش صحت‌سنجی

Table 5- Observational and estimated statistical characteristics of the models for the validation section

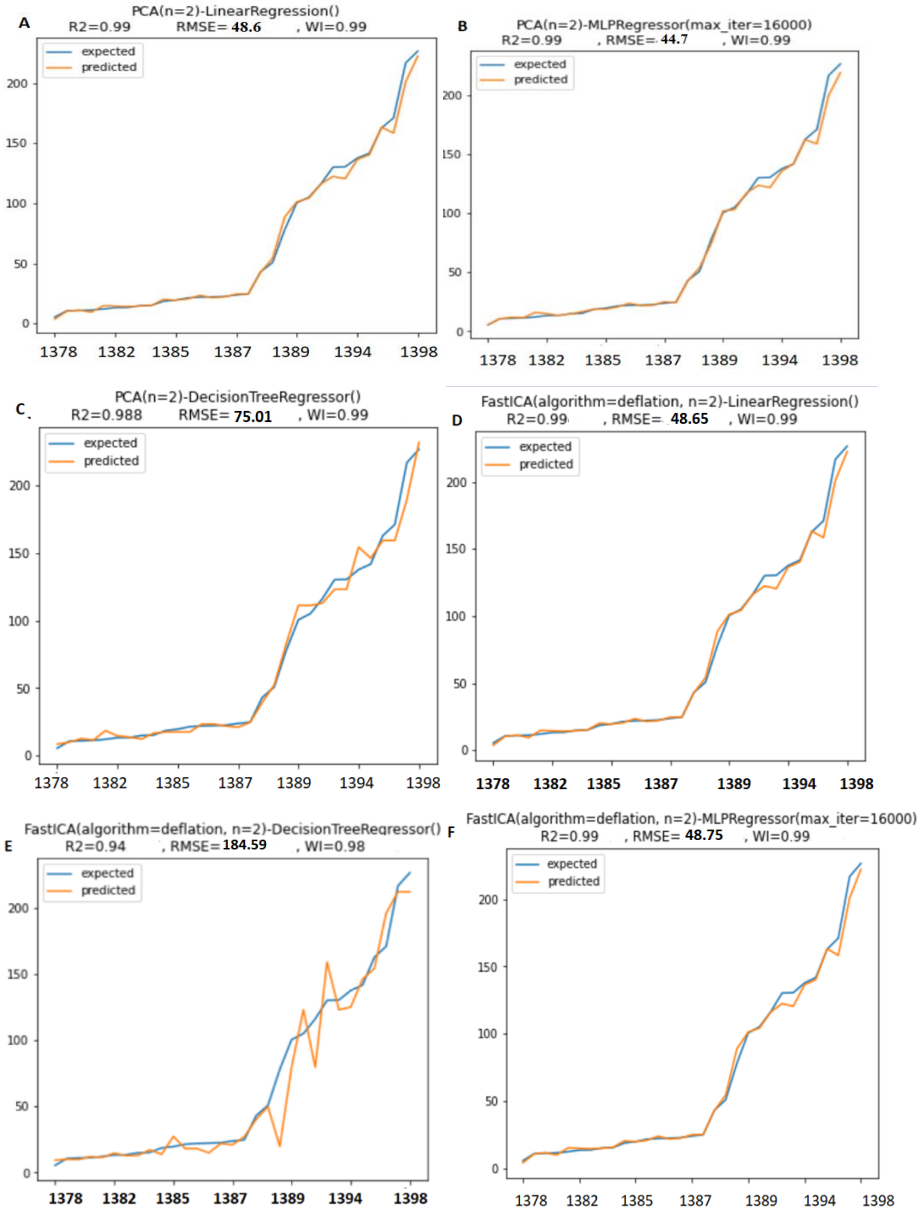
روش ارائه شده	Min	Max	Mean	Variance	Standard deviation	Skewness	Kurtosis
PCA-MLPR	5.85	225.61	66.45	4186.19	64.07	0.94	0.36
ICA-MLPR	3.858	222.61	66.45	4186.19	64.70	0.93	0.36
ICA-LR	8.604	231.74	67.20	4326.23	65.77	0.92	0.40
PCA-LR	5.592	219.16	66.09	4118.89	64.18	0.94	0.39
PCA-DTreeR	9.438	212.38	64.37	4630.08	68.05	1.07	0.30
ICA-DTreeR	3.994	222.27	66.45	4167.605	64.56	0.93	0.35
WQI	5.574	226.68	53.99	4537.167	67.36	0.98	0.25

جدول ٦- ارزیابی الگوریتم‌های PCA و ICA بر اساس مدل‌های ارائه شده برای کیفیت آب

Table 6- Evaluation of PCA and ICA algorithms based on the presented models for water quality

الگوریتم‌های مورد استفاده	معیار ارزیابی	روش کاهش بعد	
		PCA	ICA
MLP	R^2	0.995	0.993
	RMSE	44.79	48.75
	WI	0.99	0.998
LR	R^2	0.994	0.992
	RMSE	48.65	48.65
	WI	0.998	0.996
Dtree	R^2	0.987	0.920
	RMSE	75.00	184.59
	WI	0.990	0.980

شکل ۳ نشان‌دهندهٔ میزان دقت الگوریتم مورد استفاده در مقایسه با مقدار واقعی موجود برای پیش‌بینی کیفیت آب است. همان‌طور که از سیر صعودی نمودارهای شکل ۳ مشخص است با گذشت زمان شاخص کیفی آب به‌شدت کاهش پیدا کرده است. قطعا کاهش دبی منجر به افت کیفیت آب خواهد شد. در این پژوهش شاخص کیفی آب مدل‌سازی شده است. در محاسبه این شاخص از ۱۲ پارامتر (پارامتر) هیدروشیمیایی استفاده شده است. اگرچه دبی در حالت کلی در کیفیت آب مؤثر است ولی در محاسبهٔ شاخص کیفی آب در نظر گرفته نمی‌شود. به‌دلیل عبور از اراضی مختلف و پیوستن آبراهه‌های فرعی در قره‌گونئی از کیفیت آن کاسته شده است. به‌عبارت دیگر در قره‌گونئی به‌دلیل اتصال شاخه‌های فرعی از کیفیت آب کاسته شده است. کیفیت آب در اکثر مواقع در فصل تابستان در بدترین حالت خود قرار گرفته است که بررسی‌ها نشان داد کاهش میزان دبی در این موضوع دخیل بوده است برای درک بهتر نتایج الگوریتم‌های مختلف، نمودار تیلور برای ایستگاه رسم شده است. هم‌چنین، نشان‌گر نتایج حاصل از استفاده از الگوریتم‌های مختلف یادگیری ماشین است.

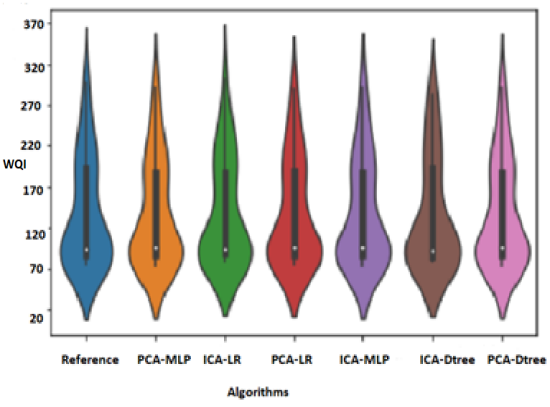


شکل ۳ – مقایسهٔ نتایج الگوریتم‌ها در پیش‌بینی کیفیت آب

Figure 3- Comparing the results of algorithms in predicting water quality

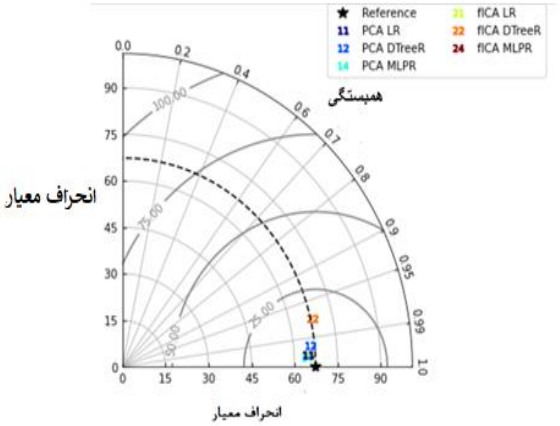
همان‌طور که از شکل ۳-ه یعنی E مشخص است روش درخت تصمیم‌گیری نتیجهٔ قابل‌قبولی ارائه نکرده و توانایی کافی برای پیش‌بینی ضریب کیفی آب را نداشته است. شکل ۳-ب (یعنی B) نشان‌دهندهٔ عملکرد بسیار مناسب روش پرسپترون چندلایه در مقایسه با سایر روش‌ها است. همان طورکه از سیر صعودی نمودارهای شکل فوق مشخص است با گذشت زمان شاخص کیفی آب به‌شدت کاهش پیدا کرده است. به‌دلیل عبور از اراضی مختلف و پیوستن آبراهه‌های فرعی در قره‌گونئی از کیفیت آن کاسته شده است. به‌عبارت دیگر در قره‌گونئی به‌دلیل اتصال شاخه‌های فرعی از کیفیت آب کاسته شده است. کیفیت آب در اکثر مواقع در فصل تابستان در بدترین حالت خود قرار گرفته است که بررسی‌ها نشان داد کاهش میزان دبی در این موضوع دخیل بوده است برای درک بهتر نتایج الگوریتم‌های مختلف، نمودار تیلور برای ایستگاه رسم شده است.

برای نمایش بهتر عملکرد مدل‌ها، نمودار تیلور و نمودار ویلونی در شکل‌های ۴ و ۵ ترسیم شده است. نمودار ویلونی نیز بیان‌گر نزدیک و قابل‌قبول بودن نتایج مدل‌ها معرفی شده است (شکل ۴). با توجه به شکل ۵ بهترین عملکرد، مربوط به الگوریتم PCA برای کاهش ابعاد و استفاده از الگوریتم MLP برای مدل‌سازی مقدار کیفی شاخص آب است. زیرا مقدار انحراف استاندارد داده‌های مشاهداتی و انحراف استاندارد داده‌های برآورد شده از مدل، نزدیک به هم بوده و نتایج قابل‌قبولی ارائه کرده است.



شکل ٤- نمودار ویلونی مقایسهٔ مدل‌های مورد استفاده در ایستگاه قره گونئی

Figure 4 - Villoni diagram comparing the models used in Qara Gunei station



شکل۵- نمودار تیلور برای ایستگاه قره‌گونئی

Figure 5- Taylor diagram for Qara Gunei station

برای آزمون پایایی مدل پیشنهادی و تعیین اعتبار مدل، نمودار تیلور توسط بسیاری از پژوهش‌گران توصیه شده است و معمولاً از آن استفاده می‌شود. از شکل ۵ می‌توان رابطهٔ بین همبستگی و انحراف استاندارد برای مدل‌سازی شاخص کیفی آب واقعی و پیش‌بینی شده در سه مدل را مشاهده کرد. مشاهده می‌شود که شبکهٔ عصبی پرسپترون چندلایه از تمام مدل‌های دیگر که در آن توزیع انحراف استاندارد برای داده‌های پیش‌بینی شده نزدیک به واقعی است بهتر عمل می‌کند نشان می‌دهد. شکل ۴ نشان‌گر نمودار ویلونی است که تمامی روش‌های اشاره شده به همراه روش‌های کاهش ابعاد با نتیجهٔ نهایی مقایسه شده است. نمودار تیلور به دو صورت نیم‌دایره نمایش همبستگی منفی و مثبت (و ربع‌دایره فقط نمایش همبستگی مثبت) ارائه می‌شود که در هر دو صورت، مقادیر ضریب همبستگی به‌صورت شعاع دایره روی قوس آن، مقادیر انحراف معیار به‌صورت دوایر متحدالمرکز نسبت به مرکز دایره و مقادیر RMSD به‌صورت دوایر متحدالمرکز نسبت به نقطهٔ مرجع دایرهٔ توخالی روی محور افقی ترسیم می‌شود. نقطهٔ مرجع کیفیت را براساس انحراف معیار سری زمانی آن نشان می دهد زیرا مقدار RMSD و ضریب تعیین سری زمانی در مقایسه با خودش به‌ترتیب صفر و یک خواهد بود، لذا موقعیت آن روی محور افقی براساس مقدار انحراف معیار تعیین خواهد شد. روش ارزیابی در این نمودار به این صورت است که داده و انحراف معیار سری زمانی روی نمودار ترسیم می‌شود که موقعیت مکانی آن روی نمودار به نقطهٔ مرجع نزدیک‌تر باشد، با دقت بیش‌تری مقدار شاخص کیفی آب را برآورد می‌کند.

بنابراین، مناسب‌تر خواهد بود همان‌طور که از نمودار مشخص است روش MLP به‌وسیلهٔ روش کاهش بعد PCA توانسته بیش‌ترین دقت را نشان دهد. این آمارها باهم، خلاصه‌ای سریع از درجهٔ مطابقت الگو را ارائه می‌دهند و می‌توان میزان دقت یک مدل شبیه‌سازی سیستم طبیعی را اندازه‌گیری کرد. نتایج حاصل از نمودار تیلور (شکل ۵) نیز با نتایج جدول پارامترهای خطا مطابقت داشته و هر دو برتر بودن پرسپترون چندلایه نسبت به سایر روش‌های یادگیری ماشین را نشان می‌دهد.

٤- نتیجه‌گیری

در پژوهش حاضر برای مدل‌سازی شاخص کیفی آب در مرحلهٔ اول، به‌منظور پیدا کردن مؤثرترین پارامترها و پارامتر وابسته به تابع هدف مطالعهٔ مورد نظر از روش‌های مختلف کاهش ابعاد از روش‌های ICA و PCA استفاده شد. در مرحلهٔ دوم از روش‌های یادگیری ماشین مانند درخت تصمیم‌گیری، رگرسیون خطی و پرسپترون چندلایه استفاده شد. در روش استفاده شده توسط Tripathi and Singal (2019)، با استفاده از روش تحلیل مؤلفهٔ اصلی تعداد پارامترهای مورد نیاز برای محاسبهٔ شاخص کیفی را از ۲۸ به ۹ کاهش دادند و با تعداد ۹ پارامتر به محاسبهٔ شاخص کیفی آب پرداختند. این در حالی است که روش ارائه شده به بررسی دو روش PCA و ICA ابعاد مسأله را از ۱۲ بعد به ۲ بعد رسانده است. مقایسهٔ نتایج مدل‌ها با استفاده از معیارهای ارزیابی عددی و گرافیکی متفاوتی از جمله R^2، RMSE و ضریب ویلموت اصلاح شده به‌عنوان معیارهای عددی و نمودار تیلور به‌عنوان معیارهای گرافیکی به‌کار گرفته شد. نتایج نشان می‌دهد روش PCA می‌تواند به ما در بهبود عملکرد با هزینهٔ کم و دقت بالا کمک کند. چرا که الگوریتم PCA می‌تواند به کاهش خطا در داده‌ها، انتخاب ویژگی و تولید ویژگی‌های مستقل و نامرتبط از داده‌ها کمک کند. همین امر باعث شده است بتواند نسبت به روش ICA عملکرد بهتری داشته باشد. نتایج نشان می‌دهد روش‌های پرسپترون چندلایه، درخت تصمیم و رگرسیون لجستیک به‌ترتیب با دقت بالا شاخص کیفی آب را انجام بدهند. در این پژوهش برای اولین‌بار با استفاده از الگوریتم کاهش بعد ICA ضمن کاهش ابعاد مسأله، شاخص کیفی آب با دقت بالای ۹۰ درصد پیش‌بینی شد. محدودیت اساسی این پژوهش استفاده از داده‌های نسبتاً کوتاه‌مدت (۲۱ سال) و یک ایستگاه است. این محدودیت باعث می‌شود که نتایج به‌دست آمده در این پژوهش قابل استفاده در سایر رودخانه‌ها نباشد. با توجه به نتایج مذکور، پیشنهاد می‌شود روش‌های مورد استفاده در مطالعهٔ حاضر در حوضه‌های آبریز با اقلیم‌های مختلف مورد بررسی قرار گرفته و برای هر اقلیم، بهترین روش مشخص شود. با توجه به زیاد بودن هزینه‌های ثبت داده‌ها و تحلیل‌های آزمایشگاهی با استفاده از روش‌های ارائه شده در این پژوهش می‌توان صرفا با دو پارامتر هیدروشیمیایی نسبت به برآورد کیفی آب اقدام نمود. این موضوع به عنوان اساسی‌ترین مزیت این روش مطرح بوده که می‌تواند توسط کارشناسان شرکت‌های آب منطقه‌ای به‌عنوان یک روش دقیق و کم‌هزینه استفاده شود.

منابع

دزفولی، دنیا، موغاری، سید محمدحسین، ابراهیمی، کیومرث، و عراقی‌نژاد، شهاب (۱۳۹۶). تعیین طبقه‌بندی کیفی آب بر اساس حداقل پارامترهای کیفی (مطالعه موردی: رودخانه کارون). *محیط زیست طبیعی*، ۷۰(۳)، ۵۸۳-۵۹۵. doi: 10.22059/JNE.2017.213338.1224

خلیلی، رضا، منتصری، حسین، متقی، حامد، و جلیلی، محمدباقر (۱۴۰۰). ارزیابی کیفیت آب رودخانه تالار استان مازندران با استفاده از ترکیب شاخص‌های کیفیت آب و مدل‌سازی چندپارامتره. *مدل‌سازی و مدیریت آب و خاک*، ۱(۴)، ۳۰-۴۷. doi:10.22098/MMWS.2021.9322.1033

سلیمان‌پور، سیدمسعود، مصباح، سید حمید، و هدایتی، بهرام (۱۳۹۷). کاربرد تکنیک داده‌کاوی درخت تصمیم CART در تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی (مطالعه موردی: دشت کازرون استان فارس). سلامت و محیط زیست، ۱۱(۱)، ۱-۱۴. http://ijhe.tums.ac.ir/article-1-5881-en.html

References

Ajayram, K.A., Jegadeeshwaran, R., Sakthivel, G., Sivakumar, R., Patange, A.D. (2021). Condition monitoring of carbide and non-carbide coated tool insert using decision tree and random tree – A statistical learning. Materials Today: Proceedings, doi:10.1016/j.matpr.2021.02.065.

Al-Mukhtar, M., & Al-Yaseen, F. (2019). Modeling water quality parameters using data-driven models, a case study Abu-Ziriq marsh in south of Iraq. *Hydrology*, *6*(1), 24. doi:10.3390/ hydrology6010024

Boyacioglu, H. (2007). Development of a water quality index based on a European classification scheme. *Water SA*, *33*(1). doi: 10.4314/wsa. v33i1.47882

Bailey, D., & Solomon, G. (2004). Pollution prevention at ports: clearing the air. *Environmental Impact Assessment review*, *24*(7-8), 749-774. doi:10.1016/j.eiar.2004.06.005

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., & Ren, H. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research*, 171, 115454. doi:10.1016/j.watres. 2019.115454.

Coutsias, E.A., Seok, C., & Dill, K.A. (2004). Using quaternions to calculate RMSD. *Journal of computational chemistry*, *25*(15), 1849-1857. doi:10.1002/jcc.20110

Daffertshofer, A., Lamoth, C.J., Meijer, O.G., & Beek, P.J. (2004). PCA in studying coordination and variability: a tutorial. *Clinical biomechanics*, *19*(4), 415-428. doi:10.1016/ j.clinbiomech.2004.01.005

Denil, M., Matheson, D., de Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 32(1), 665-673. doi:10.48550/ arXiv.1310.1415

Dezfooli, D., Mooghari, S.M.H., Ebrahimi, K., & Araghinejad, S. (2017). Water quality classification based on minimum qualitative parameter (Case Study: Karun River). Journal Of Natural Environment, *70*(3), 583-595. https://sid.ir/paper/195087/en. [In Persian]

Gorde, S.P., & Jadhav, M.V. (2013). Assessment of water quality parameters: a review. J*ournal of Engineering Research and Applications*, *3*(6), 2029-2035. https://www.ijera.com/papers/Vol3_ issue6/LV3620292035.pdf

Hintze, J.L. & Nelson, R.D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*, 181-184. doi: 10.2307/268547

Islam Khan, D.S., Islam, N., Uddin, J., Islam, S., Nasir, M.K. (2021). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences*, 34(8), 4773-4781. doi: 10.1016/ j.jksuci.2021.06.003.

Icaga, Y. (2007). Fuzzy evaluation of water quality classification. *Ecological Indicators*, *7*(3), 710-718. doi:10.1016/j.ecolind.2006.08.002

Jie, Z., Xiaoli, L., & Juntao, L. (2016). Fresh food distribution center storage allocation strategy analysis based on optimized entry-item-quantity-ABC. *International Journal on Data Science Technology*, 36-40. doi: 10.11648/j.ijdst.20160 203.11

Johnson, O., Akinola, S., Aboyeji, O., Adedeji, A. (2021). Comparison between fuzzy logic and water quality index methods: A case of water quality assessment in Ikare community, Southwestern Nigeria. *Environmental Challenges*, 3, 1-10. doi:10.1016/j.envc.2021.100038.

Kalmegh, S. (2015). Analysis of WEKA data mining algorithm REPTree, simple cart and randomtree for classification of Indian News. *International Journal of Innovative Science, Engineering & Technology*, 2, 438-446.

Khoi, D.N., Quan, N.T., Linh, D.Q., Nhi, P.T.T., Thuy, N.T.D. (2022). Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. *Water*, *14*. doi: 10.3390/w14101552.

Khalili, R., Montaseri, H., Motaghi, H., & Jalili, M. B. (2021). Water quality assessment of the Talar River in Mazandaran Province based on a combination of water quality indicators and multivariate modeling. *Water and Soil Management and Modelling*, *1*(4), 30-47. doi: 10.22098/mmws.2021.9322.1033 [In Persian]

La Valley, M.P. (2008). Logistic regression. *Circulation*, *117*(18), 23952399. doi:10.1161/ CIRCULATIONAHA.106.682658

Massoud, M.A. (2012). Assessment of water quality along a recreational section of the Damour River in Lebanon using the water quality index. *Environmental Monitoring and Assessment*, *184*, 4151-4160. doi:10.1007 /s10661-011-2251-zt

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta numerica*, *8*, 143-195. doi:10.1017/S0962492900002919.

Soleimanpour, S.M., Mesbah, S.H., Hedayati, B. (2018). Application of CART decision tree data mining to determine the most effective drinking water quality factors (case study: Kazeroon plain, Fars province). *Iranian Journal of Health and Environment*, *11*(1), 1-14. http://ijhe.tums.ac.ir/article-1-5881-en.html. [In Persian]

Sattari, M.T., Feizi, H., Colac, M., Ozturk, A., Ozturk, F., & Apaydin, H. (2021). Surface water quality classification using data mining approaches Irrigation along the Aladag River. *Irrigation and Drainage*, *70*(5), 1227–1246. doi:10. 1002/ird.2594.

Tripathi, M., Singal, S. (2019). Use of principal component analysis for parameter selection for development of a novel water quality index: A case study of river Ganga India. *Ecological Indicators*, *96*, 430-436. doi:10.1016/j.ecolind.2018.09.025.

Taylor, K.E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(7), 7183-7192. doi:10.1029/2000JD900719

Willmott, C.J., Robeson, S.M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, *32*(13), 2088-2094. doi:10.1002/joc.2419

World Health Organization. (2010). Hardness in drinking-water: background document for development of WHO guidelines for drinking-water quality (No. WHO/HSE/WSH/10.01/10).

Yusri, H., Ab Rahim, A., Hassan, S., Halim, I., & Abdullah, N. (2022). Water quality classification using SVM and XGBoost method, IEEE 13th Control and System Graduate Research Colloquium (ICSGRC). 231-236. doi: 10.1109/ICSGRC55096.2022.9845143.

Zeinalzadeh, K., & Rezaei, E. (2017). Determining spatial and temporal changes of surface water quality using principal component analysis. *Journal of Hydrology: Regional Studies*, *13*, 1-10. doi:10.1016/j.ejrh.2017.07.002