Journal of Water and Soil
https://jsw.um.ac.ir

Research Article
Vol. 36, No. 6, Feb.-Mar., 2023, p. 695-709

Numerical Estimation of Drinking Water Quality Index Using Tree Methods and Combined Wavelet Approaches and Principal Component Analysis

M.T. Sattari[1]*, S. Javidan[2]

Received: 26-08-2022
Revised: 08-11-2022
Accepted: 28-11-2022
Available Online: 22-02-2023

How to cite this article:
Sattari, M.T., & Javidan, S. (2023). Numerical Estimation of Drinking Water Quality Index Using Tree Methods and Combined Wavelet Approaches and Principal Component Analysis. *Journal of Water and Soil* *36*(6): 695-709. (In Persian with English abstract)
DOI: 10.22067/jsw.2022.78452.1196

Introduction

Surface and underground waters are one of the world's most important problems and environmental concerns. In the last few decades, due to the rapid growth of the population, the water needs have increased, followed by the input load to the water. In order to classify the quality of underground water and water level according to the type of consumption, there are many methods, one of the most used methods is the use of quality indicators. Considering the facilities available in water quality monitoring stations and the need to save time and money, using alternative methods of modern data mining methods can be good for predicting and classifying water quality. The process of water extraction for domestic use, agricultural production, mineral industrial production, electricity production, and ester methods can lead to the deterioration of water quality and quantity, which affects the aquatic ecosystem, that is, the set of organisms that live and interact. Therefore, it is very important to evaluate the quality of surface water in water-environmental management and in monitoring the concentration of pollutants in rivers. The aim of the current research was to estimate the numerical values of the drinking water quality index (WQI) using the tree method and investigate the effect of wavelet transformation, the Bagging method, and principal component analysis.

Materials and Methods

In this research, to calculate the WQI index from the quality parameters of the Bagh Kalaye hydrometric station including total hardness (TH), alkalinity (pH), electrical conductivity (EC), total dissolved solids (TDS), calcium (Ca), sodium (Na), Magnesium (Mg), potassium (K), chlorine (Cl), carbonate (CO_3), bicarbonate (HCO_3) and sulfate (SO_4) were used in the statistical period of 23 years (1998-2020). Quantitative values calculated with the WQI index were considered as target outputs. By using the relief and correlation method, the types of input combinations were determined. The random tree method was used to estimate the numerical values of the WQI index. Then, the capability of the combined approach of wavelet, principal component analysis, and Bagging method with random tree base algorithm was evaluated. To compare the values obtained from the data mining methods with the values calculated from the WQI index, the evaluation criteria of correlation coefficient (R), root mean square error (RMSE), mean absolute error (MAE), and modified Wilmot coefficient (Dr) were used.

Results and Discussion

The use of the wavelet transform method and the Bagging method has improved the modeling results. Considering that the Bagging classification method with the random tree base algorithm is a combination of the results of several random trees, so using this method has increased the accuracy of the RT model. So, in general, it was concluded that the use of wavelet transformation and classification methods increases accuracy and reduces errors. The best scenario with the highest accuracy and the lowest error was related to scenario 10 of the

1 and 2- Associate Professor and M.Sc Student, Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran, respectively.
(*- Corresponding Author Email: mtsattar@tabrizu.ac.ir)

W-B-RT model with Total Hardness, Electrical Conductivity, Total Dissolved Solid, Sulphate, Calcium, Bicarbonate, Magnesium, Chlorine, Sodium, and potassium parameters. The results showed that the effect impact of pH in estimating the numerical value of the WQI index is considered lower than other parameters. When the principal component analysis method was used, by reducing the value of the eigenvalue from F1 to F12, the value of the factor also decreased; As a result,so F1, F2, and F3 factors were selected as the basic components. Considering 3 main factors, modeling was done employed and R=0.98, RMSE=2.17, MAE=1.52, and Dr=0.97 were obtained. In general, the results showed that the PCA method, despite reducing the dimension of the input vectors and simplifying it, can improve the accuracy and speed of the model and is introduced as the best method for estimating the numerical value of the WQI index.

Conclusion

The results obtained from the present research showed that the use of wavelet transform, Bagging and PCA methods had a positive effect on improving the results and increasing higherthe accuracy. In estimating the numerical values of WQI index, PCA-B-RT method considering 3 main factors, with correlation coefficient equal to 0.98, root mean square error equal to 2.17, average absolute value error equal to 1.52 and tThe modified Wilmot coefficient equal to 0.97 had the highest accuracy. Considering that all the methods used in the estimation of quantitative values had acceptable accuracy, therefore, in case of lack of data and lack of access to all chemical parameters, it is possible to obtain appropriate and acceptable results by using a limited number of parameters and data mining methods achieved.

Keywords: Bagging preprocessing approach, Modified Wilmot coefficient, Principal component analysis, Relief Algorithm, Wavelet transform

Journal of Water and Soil
https://jsw.um.ac.ir

مقاله پژوهشی

جلد ۳۶، شماره ۶، بهمن-اسفند ۱۴۰۱، ص. ۷۰۹-۶۹۵

تخمین عددی شاخص کیفی آب شرب با استفاده از روش‌های درختی و رویکردهای ترکیبی موجک و تحلیل مؤلفه اصلی

محمدتقی ستاری[1*]- سحر جاویدان[2]

تاریخ دریافت: ۱۴۰۱/۰۶/۰۴

تاریخ پذیرش: ۱۴۰۱/۰۹/۰۷

چکیده

آگاهی از کیفیت آب، یکی از نیازهای مهم در برنامه‌ریزی، توسعه و حفاظت از منابع آب برای مصارف مختلف از جمله شرب به شمار می‌رود. استفاده از روش‌های مدرن داده‌کاوی، می‌تواند رویکرد مناسبی برای پیش‌بینی و طبقه‌بندی کیفیت آب باشد. در پژوهش حاضر، برای محاسبه شاخص کیفی آب شرب از پارامترهای شیمیایی شامل سختی کل، قلیائیت، هدایت الکتریکی، کل مواد جامد محلول، کلسیم، سدیم، منیزیم، پتاسیم، کلر، کربنات، بی‌کربنات و سولفات ایستگاه هیدرومتری باغ کلایه استان قزوین، در دوره آماری ۲۳ ساله (۱۹۹۸-۲۰۲۰) استفاده شد. روش درخت تصادفی برای تخمین و مدل‌سازی مقادیر عددی شاخص کیفی آب شرب براساس ترکیب‌های مختلفی از پارامترهای شیمیایی به کار برده شد. ماتریس همبستگی و الگوریتم رلیف، مبنای انتخاب ترکیب‌های مختلفی از پارامترهای شیمیایی به‌عنوان ورودی روش‌های داده‌کاوی در قالب سناریوهای مختلف در نظر گرفته شدند. در جهت بهبود نتایج تخمین عددی شاخص کیفی آب شرب، از رویکردهای تبدیل موجک، دسته‌بندی مدل‌ها و تحلیل مؤلفه اصلی استفاده شد. بررسی نتایج نشان داد که ترکیب ۳ روش تحلیل مؤلفه اصلی (با در نظر گرفتن ۳ عامل اصلی)، رویکرد پیش‌پردازش Bagging و درخت تصادفی، با ضریب همبستگی برابر با ۰/۹۸، ریشه میانگین مربعات خطا برابر با ۲/۱۷، میانگین خطای قدر مطلق برابر با ۱/۵۲ و ضریب ویلموت اصلاح شده برابر با ۰/۹۷ می‌تواند دقت بالایی در تخمین مقادیر عددی شاخص کیفی آب شرب داشته باشد. براساس نتایج کلی به دست آمده، در صورت کمبود نمونه‌های آزمایشگاهی و یا عدم دسترسی به تمام پارامترهای شیمیایی، روش‌های معرفی شده در این مطالعه، به علت دقت بالا جهت تخمین شاخص کیفی آب شرب قابل توصیه خواهند بود.

واژه‌های کلیدی: الگوریتم رلیف، تبدیل موجک، تحلیل مؤلفه اصلی، رویکرد پیش‌پردازش Bagging، ضریب ویلموت اصلاح شده

مقدمه

آلودگی آب‌های سطحی و زیرزمینی از مهم‌ترین معضلات جهان و نگرانی‌های زیست محیطی محسوب می‌شود. در چند دهه اخیر به‌علت رشد سریع جمعیت، نیازهای آبی و به دنبال آن بار آلودگی ورودی به منابع آب افزایش یافته‌است. جهت طبقه‌بندی کیفیت آب با توجه به نوع مصرف، روش‌های متعددی وجود دارد که یکی از روش‌های پرکاربرد، استفاده از شاخص‌های کیفی است. با توجه به کمبود امکانات در تمامی ایستگاه‌های رصد کیفیت آب و نیاز به صرفه جویی در زمان و هزینه، استفاده از روش‌های جایگزین مانند روش‌های مدرن داده‌کاوی می‌تواند رویکرد مناسبی برای پیش‌بینی و طبقه‌بندی کیفیت آب باشد (Kavita and Jagdish, 2012). فرآیند برداشت آب برای مصارف خانگی، تولید کشاورزی، تولید صنعتی- معدنی، تولید برق و غیره می‌تواند منجر به بدتر شدن کیفیت آب شده و بر اکوسیستم آبی تأثیر منفی بگذارد. بنابراین ارزیابی کیفیت آب‌های

۱ و ۲- به‌ترتیب دانشیار و دانشجوی کارشناسی ارشد منابع آب، گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران

(*- نویسنده مسئول: (Email: mtsattar@tabrizu.ac.ir

DOI: 10.22067/jsw.2022.78452.1196

سطحی در مدیریت آبی-محیطی و در پایش غلظت آلاینده‌ها در رودخانه‌ها بسیار حائز اهمیت است. بررسی منابع نشان می‌دهد؛ تحقیقات زیادی در زمینه تخمین کیفیت آب‌های سطحی و پیش‌بینی شاخص کیفی آب با استفاده از روش‌های داده‌مبنا در سطح ملی و بین‌المللی انجام یافته است. سلیمان‌پور و همکاران (Soleimanpour *et al.*, 2018) برای تعیین مؤثرترین عامل کیفیت آب آشامیدنی دشت کازرون، از تکنیک داده‌کاوی درخت تصمیم CART[1] استفاده کردند. نتایج مطالعه آن‌ها 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 دانسته‌اند. باتور و مکتاو (Bature and Maktav, 2019) کیفیت آب‌های سطحی را در دریاچه گالا کشور ترکیه با استفاده از فیوژن تصاویر ماهواره‌ای بر اساس روش تحلیل مؤلفه اصلی، ارزیابی نمودند. آن‌ها پس از انجام تجزیه و تحلیل‌های لازم، نتیجه گرفتند که روش رگرسیون سطح پاسخ (RSR[2]) مبتنی بر PCA نسبت به مدل‌های داده‌کاوی MLR[3]، ANN[4] و SVM[5] برای تخمین دقیق پارامترهای کیفیت آب در دریاچه‌ها برتری دارد. آل مختار و آل یاسین (Al- Mukhtar and Al- Yaseen, 2019) با استفاده از مدل‌های داده‌محور، پارامترهای کیفیت آب را مدل‌سازی کردند. آن‌ها روش ANFIS[6] را به‌عنوان یک مدل پیش‌بینی‌کننده برای TDS[7] و EC در عراق پیشنهاد کردند. همچنین نتیجه گرفتند که نیترات، کلسیم، منیزیم، سختی کل (TH[8])، سولفات و کلرید، تأثیرگذارترین ورودی‌ها در TDS و کلسیم، منیزیم، سختی کل، سولفات و کلرید بیشترین تأثیر را روی EC دارند. حسینی و همکاران (Hosseini *et al.*, 2019) برای ارزیابی کیفی آب سطحی استان سیستان و بلوچستان، شاخص کیفیت آب (WQI) را به کار بردند. نتایج تحلیل‌های آماری انجام شده، مشخص کرد که همبستگی میان پارامترهایی مانند کلیفرم، نیترات، سولفات و کلر با شاخص کیفیت آب، زیاد بوده است. اتهمان و همکاران (Othman *et al.*, 2020) به پیش‌بینی شاخص کیفیت آب رودخانه با در نظر گرفتن حداقل تعداد متغیرهای ورودی پرداختند. نتایج، بیانگر توانایی استثنایی مدل شبکه عصبی مصنوعی برای محاسبه WQI[9] بود. همچنین آن‌ها اکسیژن محلول (DO[10]) را به‌عنوان مؤثرترین پارامتر در تعیین کیفیت آب معرفی کردند. نیهالانی و مروتی (Nihalani and Meeruty, 2020) شاخص کیفیت آب رودخانه‌های اصلی در گجرات هند را ارزیابی نمودند. نتایج مطالعه آن‌ها نشان داد که مقدار شاخص کیفیت آب برای رودخانه ماهی ۳۰ تا ۵۰، برای رودخانه سابارماتی ۴۲ تا ۶۵، برای رودخانه نارمادا ۲۸ تا ۵۲ و برای رودخانه تاپی ۳۵ تا ۷۰ است. بر اساس WQI، استنباط کردند که دلیل اصلی کاهش کیفیت آب رودخانه، تخلیه فاضلاب، پساب صنعتی و رواناب شهری بوده است. چن و همکاران (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 بر اساس داده‌های بزرگ پرداختند. آن‌ها برای پایش کیفیت آب در آینده و ارائه هشدار به موقع کیفیت آب، روش درخت تصمیم[11]، جنگل تصادفی[12] و آبشار عمیق[13] را در اولویت قرار دادند. خوی و همکاران (Khoi *et al.*, 2022) برای پیش‌بینی شاخص کیفیت آب در رودخانه لا بوونگ ویتنام، از مدل‌های یادگیری ماشین استفاده کردند. نتایج تحقیق آن‌ها نشان داد که مدل تقویت گرادیان شدید[14] (XGBoost) با R=0.989 و RMSE=0.107، عملکرد خوبی در پیش‌بینی WQI داشت. یافته‌های آن‌ها این استدلال را تقویت کرد که مدل‌های یادگیری ماشین، به ویژه XGBoost، ممکن است برای پیش‌بینی WQI با سطح بالایی از دقت، مورد استفاده قرار گیرند و مدیریت کیفیت آب را بهبود بخشند.

هدف پژوهش حاضر، محاسبه مقادیر عددی شاخص WQI با استفاده از داده‌های مربوط به پارامترهای کیفی آب ایستگاه هیدرومتری باغ کلایه در استان قزوین در دوره آماری ۲۳ ساله (۱۹۹۸-۲۰۲۰) است. مقادیر عددی این شاخص با استفاده از روش درخت تصادفی، تخمین زده شده و قابلیت رویکرد پیش‌پردازش Bagging، تبدیل موجک، تحلیل مؤلفه اصلی نیز در جهت بهبود نتایج مدل‌سازی مورد بررسی قرار گرفت.

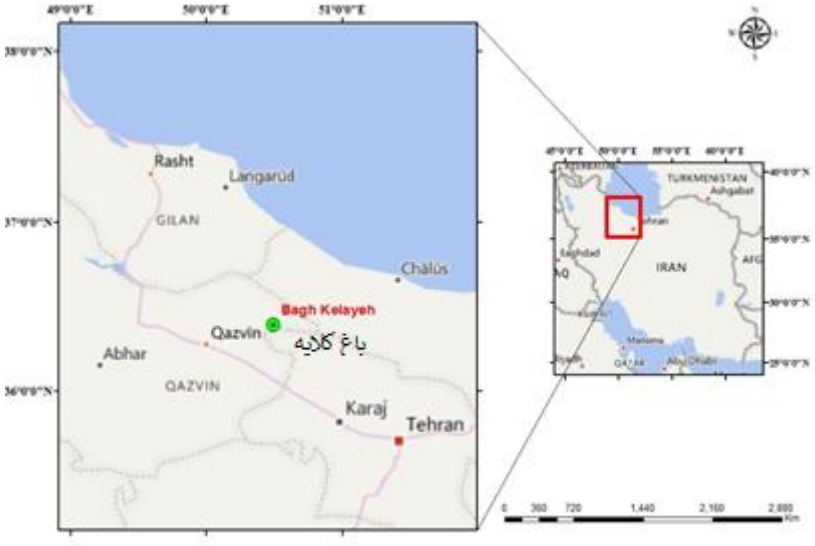
مواد و روش‌ها

منطقه مورد مطالعه و داده‌های مورد استفاده

استان قزوین در بخش شمال غربی کشور ایران واقع شده و مساحت آن حدود ۱۵۸۲۰ کیلومتر مربع است. باغ کلایه، روستایی از

1- Classification And Regression Trees
2- Response Surface Regression
3- Multiple Linear Regression
4- Artificial Neural Network
5- Support Vector Machines
6- Adaptive Neuro Fuzzy Inference System
7- Total Dissolved Solid
8- Total Hardness
9- Water Quality Index
10- Dissolved Oxygen
11- Decision Tree
12- Random Forest
13- Deep Cascade Forest
14- Extreme Gradient Boosting

توابع بخش رودبارالموت شهرستان قزوین در استان قزوین است. ایستگاه هیدرومتری باغ کلایه در عرض جغرافیایی ۳۶ درجه و ۲۳ دقیقه و ۳۸ ثانیه، طول جغرافیایی ۵۰ درجه و ۲۹ دقیقه و ۵۱ ثانیه و ارتفاع ۱۲۸۷ متر از سطح دریا واقع شده است. موقعیت مکانی ایستگاه مورد مطالعه در شکل ۱ آورده شده است.



شکل ۱- موقعیت مکانی ایستگاه مورد مطالعه

Figure 1- Location of the studied station

جدول ۱- مشخصات آماری پارامترهای مورد استفاده

Table 1- Statistical characteristics of implemented parameters

حداقل Minimum	حداکثر Maximum	میانگین Mean	واریانس Variance	ضریب تغییرات Coefficient Variation	معیارهای آماری Statistical criteria
95.00	481.50	278.15	3407.38	0.210	سختی کل (میلی‌گرم بر لیتر) TH ($mg.L^{-1}$)
4.50	8.40	7.83	0.11	0.042	قلیاییت pH
279.00	1048.00	627.94	17480.87	0.211	هدایت الکتریکی (میکروموس بر سانتی‌متر) Ec (µ mho/cm)
186.00	663.00	388.05	5983.81	0.199	کل مواد جامد محلول (میلی‌گرم بر لیتر) TDS ($mg. L^{-1}$)
0.00	159.80	73.94	380.24	0.264	کلسیم (میلی اکی‌والان بر لیتر) Ca ($meq. L^{-1}$)
0.46	60.49	17.26	82.98	0.528	سدیم (میلی اکی‌والان بر لیتر) Na ($meq. L^{-1}$)
2.76	58.20	22.18	0.9370	0.436	منیزیم (میلی اکی‌والان بر لیتر) Mg ($meq. L^{-1}$)
0.39	19.50	1.99	2.79	0.839	پتاسیم (میلی اکی‌والان بر لیتر) K ($meq. L^{-1}$)
0.00	89.60	27.36	198.61	0.515	کلر (میلی اکی‌والان بر لیتر) Cl ($meq. L^{-1}$)
0.00	33.00	0.21	4.17	9.808	کربنات (میلی اکی‌والان بر لیتر) CO_3 ($meq. L^{-1}$)
22.08	374.88	140.21	2129.19	0.329	سولفات (میلی اکی‌والان بر لیتر) SO_4 ($meq. L^{-1}$)
50.02	391.62	163.60	1565.21	0.242	بی کربنات (میلی اکی‌والان بر لینر) HCO_3 ($meq. L^{-1}$)

در پژوهش حاضر برای محاسبه شاخص WQI از پارامترهای کیفی ایستگاه هیدرومتری باغ کلایه شامل سختی کل (TH)، قلیائیت (pH)، هدایت الکتریکی (EC)، کل مواد جامد محلول (TDS)، کلسیم (Ca)، سدیم (Na)، منیزیوم (Mg)، پتاسیم (K)، کلر (Cl)، کربنات (CO_3)، بی کربنات (HCO_3) و سولفات (SO_4) در دوره آماری ۲۳ ساله (۱۹۹۸-۲۰۲۰) استفاده شد. مشخصات آماری متغیرهای مورد استفاده در جدول ۱ ارائه گردید.

مقادیر کمی محاسبه شده با شاخص WQI به‌عنوان خروجی‌های هدف مدل درخت تصمیم درنظر گرفته شدند. با استفاده از روش رلیف[1] و همبستگی، انواع ترکیب‌های ورودی مشخص گردید. این ترکیب‌ها در جدول ۲ نشان داده شده است. کیرا و رندل (Kira and Rendell, 1992) استفاده از الگوریتم رلیف را که یک الگوریتم انتخاب ویژگی برای کاهش ابعاد مسئله پیشنهاد دادند. این الگوریتم نکات قوتی دارد که می‌توان به ساده بودن اصول و عدم پیچیدگی آن، قابل حل بودن با توابع چندجمله‌ای مرتبه پایین، قابل استفاده بودن برای داده‌های پیوسته و نیاز به تعداد کم داده‌های آموزشی اشاره کرد. در یک مجموعه داده با تعداد N نمونه (داده مشاهداتی) و تعداد P ویژگی که مربوط به دو طبقه مختلف هستند، هر ویژگی باید در بازه (۰،۱) قرار گیرد. الگوریتم مذکور، m بار تکرار شده و در هر مرتبه از یک بردار وزنی متفاوت که از صفر شروع می‌گردد، استفاده می‌کند. در هر تکرار، الگوریتم مذکور بردار ویژگی X را که متعلق به یک نمونه تصادفی است و بردارهای ویژگی نزدیک‌ترین نمونه به نمونه X در طبقه مورد نظر را توسط تابع فاصله اقلیدسی[2] انتخاب می‌کند. پس از mتکرار، هر یک از عناصر بردار وزن توسط m تقسیم‌بندی می‌شوند. نتیجه این عمل این است که یک بردار مرتبط به دست می‌آید؛ چنان‌چه مقدار بردار مرتبط یک ویژگی، از آستانه تعریف شده بیشتر گردد، آن ویژگی انتخاب می‌گردد. از بین داده‌های موجود، ۷۰ ٪ برای واسنجی و ۳۰ ٪ برای صحت‌سنجی در نظر گرفته شدند.

برای برآورد مقادیر عددی شاخص WQI از روش درخت تصادفی[3] استفاده شد. سپس قابلیت رویکرد پیش‌پردازش Bagging، تبدیل موجک و تحلیل مؤلفه اصلی ارزیابی شد.

روش‌های مورد مطالعه

شاخص کیفیت آب شرب

از زمان هورتون در سال ۱۹۶۵، بسیاری از نویسندگان به دنبال تجمیع متغیرهای مختلف توصیف‌کننده وضعیت آب در یک مقدار واحد به نام شاخص کیفیت آب (WQI) بودند. این شاخص برای تسهیل مدیریت عملیاتی منابع آب و تخصیص آن‌ها برای مصارف مختلف در نظر گرفته شده است. هدف WQI طبقه‌بندی آب‌ها نسبت به ویژگی‌های بیولوژیکی، شیمیایی و فیزیکی است که کاربردهای احتمالی آن‌ها را تعریف می‌کند و تخصیص آن‌ها را مدیریت می‌کند (Khalil *et al.*, 2011). برای این منظور، متغیرهای تحلیلی باید موزون و تجمیع شوند. WQIها را می‌توان به‌عنوان مدل‌هایی از کیفیت آب در نظر گرفت.

در مقاله حاضر، شاخص کیفیت آب شرب با استفاده از فرمول‌های ۱ تا ۳ محاسبه گردید. در این فرمول‌ها w وزن مربوط به هر پارامتر با توجه به اهمیت آن در شرب و W وزن نسبی هر پارامتر، C غلظت هر پارامتر، S غلظت استاندارد هر پارامتر، q رتبه کیفی هر پارامتر و WQI نیز شاخص کیفی آب شرب می‌باشد (Singh 1992).

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

$$q_i = \left(\frac{C_i}{S_i}\right) \times 100 \quad (۲)$$

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

مقادیر WQIمحاسبه شده معمولا به پنج دسته آب عالی، خوب، بد، بسیار ضعیف ونامناسب برای آشامیدن تقسیم می‌شوند (جدول ۳):

روش درخت تصادفی

درخت تصادفی یک طبقه‌بندی‌کننده تحت نظارت است و از یک ایده جمع‌آوری برای تولید مجموعه‌ای تصادفی از داده‌ها برای ساخت درخت تصمیم استفاده می‌کند. طبقه‌بندی به این صورت عمل می‌کند که طبقه‌بندی‌کننده، درختان تصادفی بردار ویژگی ورودی را می‌گیرد، آن را با هر درخت در جنگل، طبقه‌بندی می‌کند و برچسب کلاسی را که اکثریت «رای» را دریافت کرده است، به‌عنوان خروجی به ما می‌دهد. در مورد رگرسیون، پاسخ طبقه‌بندی‌کننده، میانگین پاسخ‌ها در تمام درختان جنگل است. همه درختان با پارامترهای یکسان اما در مجموعه‌های آموزشی متفاوت آموزش داده می‌شوند (Ajayram 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 تصادفی بودن در یک درخت می‌توان به آن اشاره کرد؛ ساخت هر درخت با استفاده از یک مجموعه داده بوت استرپ یا زیر نمونه‌برداری است.

1- Relief
2- Euclidean Distance Function
3- Random Tree

جدول ۲- پارامترهای دخیل در هر سناریو و روش انتخاب سناریوها

Table 2- The parameters involved in each scenario and the method of selecting scenarios

شماره سناریو (Scenario Number)	ورودی (Input)	روش انتخاب سناریو (Scenario selection method)
1	TH	Correlation Matrix
2	TH, EC	
3	TH, EC, TDS	
4	TH, EC, TDS, SO_4	
5	TH, EC, TDS, SO_4, Ca	
6	TH, EC, TDS, SO_4, Ca, HCO_3	
7	TH, EC, TDS, SO_4, Ca, HCO_3, Mg	
8	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl	
9	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na	
10	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K	
11	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K, CO_3	
12	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K, CO_3, PH	
13	TH, K	Relief
14	TH, K, SO_4	
15	TH, K, SO_4, TDS	
16	TH, K, SO_4, TDS, EC	

جدول ۳- طبقه بندی کیفیت آب بر اساس ارزش WQI

Table 3- Water quality classification based on WQI value

دامنه WQI (WQI Range)	کلاس (Class)	نوع آب (Type of Water)
below 50 (زیر ۵۰)	I	(Excellent water) عالی
50-100	II	(Good water) خوب
100-200	III	(Poor water) ضعیف
200-300	IV	(Very poor water) خیلی ضعیف
above 300 (بالای ۳۰۰)	V	(Water unsuitable for drinking) غیر قابل شرب

(طبقه‌بندی کیفیت آب آشامیدنی)

Classification of Drinking Water Quality

به این ترتیب، هر درخت در جنگل، بر روی داده‌های کمی متفاوتی آموزش داده می‌شود، که تفاوت‌های بین درختان را معرفی می‌کند (Denil *et al*., 2014).

روش دسته‌بندی

اولین بار، بریمن در سال ۱۹۹۶ روش دسته‌بندی را ارائه داد، در این روش، چندین یادگیر پایه به‌صورت موازی به هم متصل می‌شوند تا واریانس مجموعه کاهش یابد. هر یادگیر پایه با استفاده از الگوریتم یادگیری یکسان بر روی یک نسخه بوت استرپ[1] آموزش می‌بینند. مجموعه داده بوت استرپ، مجموعه‌ای است که به‌طور تصادفی و همراه با جایگزینی تمام اطلاعات مربوط به یکی از افراد جمعیت مرجع، استخراج می‌شود. این فرآیند تا زمان یکسان‌شدن تعداد مشاهدات مجموعه جدید با جمعیت مرجع ادامه پیدا می‌کند. به‌دلیل دادن شانس انتخاب مجدد به هر نمونه از جمعیت اصلی، ممکن است برخی از نمونه‌های جمعیت مرجع، چندین بار در یک مجموعه بوت استرپ ظاهر شوند؛ اما برخی دیگر اصلا انتخاب نشوند. تعداد مشاهدات، در بوت استرپ‌ها یکسان است؛ ولی هر مجموعه بوت استرپ از دیگری متفاوت خواهد بود (Kheirabadi *et al*., 2017). خروجی‌های این یادگیرندگان پایه با رأی اکثریت (برای طبقه‌بندی) یا میانگین‌گیری (برای رگرسیون) برای به‌دست آوردن خروجی نهایی جمع می‌شوند. چنان‌چه یادگیرندگان پایه در یک گروه، دقیق و متنوع باشند، می‌توان به عملکردی بهتر و قوی‌تر دست یافت. در ادبیات علم داده و داده‌کاوی روش دسته‌بندی یا Bagging یکی از روش‌های یادگیری جمعی[2] است. منظور از روش دسته‌بندی یا Bagging میانگین‌گیری از نتیجه پیش‌بینی چندین روش طبقه‌بندی است تا دقت پیش‌بینی‌ها افزایش یابد (Breiman, 1996).

تبدیل موجک

موجک[3] دسته‌ای از توابع ریاضی، برای تجزیه سیگنال پیوسته به مؤلفه‌های فرکانسی آن است. در تبدیل موجک، سطح تفکیک هر مؤلفه برابر با مقیاس آن در نظر گرفته می‌شود. تبدیل موجک تجزیه یک تابع بر مبنای توابع موجک است. موجک‌ها که به‌عنوان موجک های مادر شناخته می‌شوند؛ نمونه‌های انتقال‌یافته و مقیاس‌شده یک تابع (موجک مادر) با طول متناهی و نوسانی شدیدا میرا هستند. تابع موجک، تابعی است که دو ویژگی مهم نوسانی بودن و کوتاه‌مدت بودن را دارد (Lau and Weng, 1995). موجک هار یکی از ساده ترین روش‌ها و موجک متعارف پرکاربرد با پشتیبانی فشرده است که

1- Bootstrap

2- Ensemble

3- Wavelet

از نظر ریاضی در بین همه خانواده موجک‌ها ساده است. موجک هار بر اساس جفت توابع تکه‌ای ثابت، ساخته شده است که به راحتی ادغام می‌شوند. ویژگی‌های شناخته شده موجک هار، عبارات صریح برای توابع موجک و مقیاس است. همچنین متعامد بودن، پشتیبانی فشرده و پراکندگی ماتریس‌ها از دیگر ویژگی‌های موجک هار در نظر گرفته می‌شود (Vishwanath *et al.*, 2021). در پژوهش حاضر از تبدیل موجک هار[1] استفاده شده است. تعداد موجک هار ایجادشده تابعی از تعداد پارامترهای ورودی مدل می‌باشد. به‌عنوان مثال وقتی تعداد پارامترهای ورودی چهار تا باشد، در آن‌صورت به تعداد ۴ مورد هار ایجاد می‌شود؛ اما وقتی تعداد پارامتر ورودی ۵ تا باشد، تعداد هار ایجادشده برابر ۸ خواهد بود. به عبارت دیگر تعداد موجک هار، توانی از ۲ در نظر گرفته می‌شود (Haar, 1910).

روش تحلیل مؤلفه اصلی

تحلیل مؤلفه اصلی از جمله روش‌های آماری چند پارامتریست که با برقراری یک ارتباط خطی بین ویژگی‌های متغیرهای ورودی اولیه مدل، از حجم اطلاعات ورودی کاسته و مؤثرترین بردارهای با ضریب همبستگی صفر را انتخاب می‌کند. یک مؤلفه اصلی را می‌توان به صورت زیر نوشت.

$$\begin{aligned} PC_1 &= a_1^1 A_1 + a_1^2 D_1 + a_1^n X_n \\ PC_2 &= a_2^1 A_1 + a_2^2 D_1 + a_2^n X_n \\ &\vdots \\ PC_n &= a_n^1 A_1 + a_n^2 D_1 + a_n^n X_n \end{aligned} \qquad (۴)$$

در روابط فوق پارامترهای PC_1 تا PC_n تعداد مؤلفه‌های اصلی، a_i^j ضریب i اُمین مؤلفه اصلی و j اُمین متغیر را نشان می‌دهند (Abdi and Williams, 2010).

در تکنیک PCA با استفاده از یک تبدیل خطی، داده‌ها از داده های چندبعدی به مختصات دیگر منتقل می‌شوند. این انتقال بر اساس حداکثر واریانس و حداقل ارتباط انجام خواهد شد. بدین منظور ابتدا ماتریس کواریانس داده‌ها تشکیل شده و سپس مقادیر ویژه و بردارهای ویژه ماتریس استخراج گردیده و مرتب می‌شوند. سرانجام، چند بردار ویژه که بیشترین میزان مقادیر ویژه را دارند، نگه داشته شده و بقیه حذف می‌شوند. کاهش داده سبب سادگی مدل پیشگو شده و زمان پردازش داده‌ها نیز کمتر خواهد شد (Mat Nawi *et al.*, 2013).

1- Haar

معیارهای ارزیابی

برای مقایسه مقادیر به دست آمده از روش‌های داده‌کاوی با مقادیر محاسبه شده از شاخص WQI از معیارهای ارزیابی ضریب همبستگی (R)، ریشه میانگین مربعات خطا[2] (RMSE)، میانگین خطای مطلق[3] (MAE) و ضریب ویلموت اصلاح شده[4] (Dr) استفاده شد. فرمول‌های آماره‌های فوق به ترتیب در روابط (۵) تا (۸) ارائه گردیده است:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 . \sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad (۵)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \qquad (۶)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{N}\left|y_i - x_i\right| \qquad (۷)$$

$$Dr = \begin{cases} when \sum_{i=1}^{N}\left|y_i - x_i\right| > C\sum_{i=1}^{N}\left|x_i - \bar{x}\right| \rightarrow \frac{C\sum_{i=1}^{N}\left|x_i - \bar{x}\right|}{\sum_{i=1}^{N}\left|y_i - x_i\right|} - 1 \\ when \sum_{i=1}^{N}\left|y_i - x_i\right| \le C\sum_{i=1}^{N}\left|x_i - \bar{x}\right| \rightarrow 1 - \frac{\sum_{i=1}^{N}\left|y_i - x_i\right|}{C\sum_{i=1}^{N}\left|x_i - \bar{x}\right|} \end{cases} \qquad (۸)$$

در روابط فوق، y_i مقدار برآورد شده از مدل، x_i مقدار محاسبه شده از شاخص کیفی آب و N تعداد داده‌ها می‌باشند.

با استفاده از نرم‌افزار Visio 2016 روندنمای مراحل انجام تحقیق ترسیم و در شکل ۲ نشان داده شد.

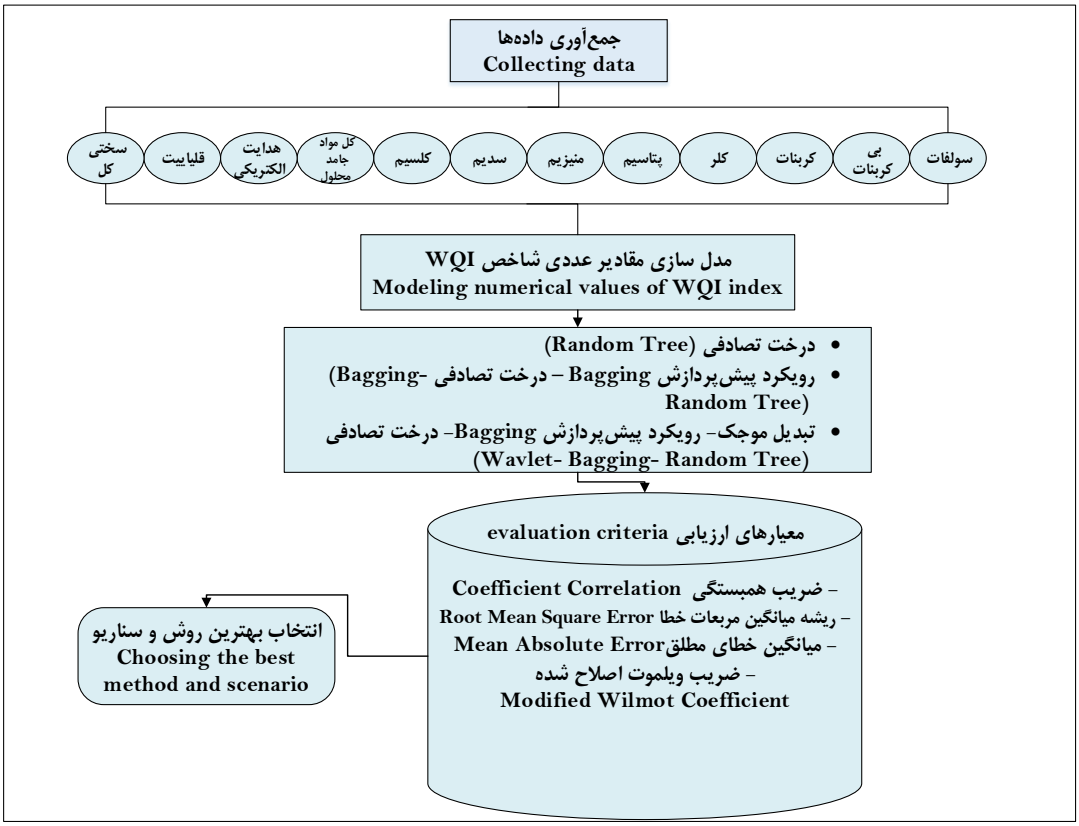
نتایج و بحث

در پژوهش حاضر برای برآورد مقادیر کمی WQI، از پارامترهای TH، pH، EC، TDS، Ca، Na، Mg، K، Cl، CO_3، HCO_3 و SO_4 ایستگاه هیدرومتری باغ کلایه در دوره آماری ۲۳ ساله استفاده شد. روش رلیف و همبستگی برای انتخاب ترکیب‌های ورودی مختلف به کار برده شد. مدل‌سازی برای برآورد مقادیر کمی WQI با روش درخت تصادفی انجام شد، سپس کارایی رویکرد پیش‌پردازش

2- Root Mean Square Error
3- Mean Absolute Error
4- Modified Wilmot Coefficient

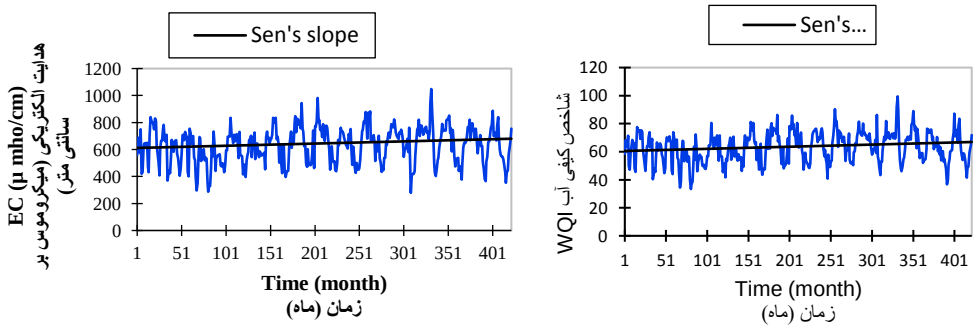
Bagging، تبدیل موجک، و PCA در جهت بهبود نتایج مدل‌سازی، بررسی شد. از بین ۴۲۲ نمونه مشاهداتی، ۷۰ درصد (۲۹۵ نمونه) برای آموزش و ۳۰ درصد (۱۲۷ نمونه) برای آزمون در نظر گرفته شد.

بررسی روند مقادیر عددی شاخص WQI، نشان داد که این شاخص، روندی افزایشی داشته و در گذر زمان با افزایش مقدار عددی آن، کیفیت آب شرب ایستگاه مورد مطالعه کمتر شده است. پس از بررسی پارامترهای تاثیرگذار در کیفیت آب شرب، مشخص گردید که مقدار هدایت الکتریکی باغ کلایه نیز روند افزایشی داشته و در نتیجه میزان شوری آن به مرور زمان، افزایش یافته است. بنابراین یکی از دلایل کاهش کیفیت آب در این ایستگاه را می‌توان افزایش شوری آب در نظر گرفت. جهت بررسی چگونگی نوسانات توام شاخص کیفی آب با مقدار هدایت الکتریکی، شاخص شیب Sen در محیط نرم‌افزار اکسل استت محاسبه و شکل ۳ رسم گردید. نتایج نشان داد که هر دو پارامتر از روند نسبی افزایشی برخوردار می‌باشند و این موضوع حکایت از کاهش نسبی کیفیت آب با هدف شرب در منطقه مورد مطالعه دارد.



شکل ۲- شمای کلی مراحل انجام تحقیق

Figure 2- Outline of the research processes



شکل ۳- نمودار سری زمانی هدایت الکتریکی و شاخص WQI

Figure 3- Time series chart of electrical conductivity and WQI index

جدول ٤- معیارهای ارزیابی برای برآورد مقادیر کمی WQI در بخش آزمون

Table 4- Evaluation criteria for estimating quantitative WQI values in the test section

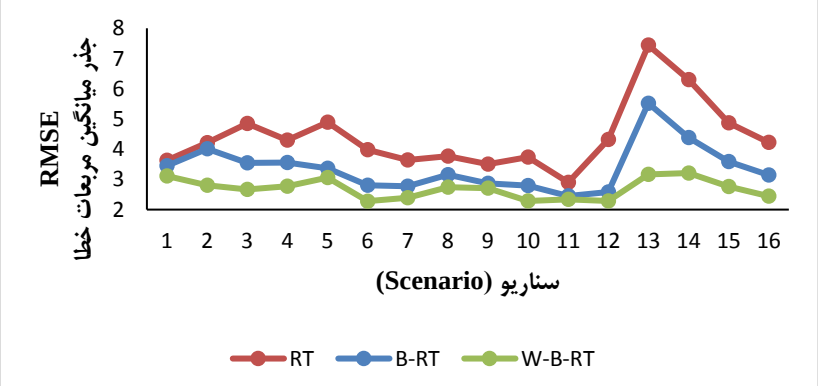
Scenario	Method											
	RT				B-RT				W-B-RT			
	R	RMSE	MAE	Dr	R	RMSE	MAE	Dr	R	RMSE	MAE	Dr
1	0.94	3.63	2.65	0.90	0.95	3.45	2.54	0.91	0.95	3.11	2.37	0.92
2	0.92	4.21	2.89	0.89	0.93	4.01	2.66	0.90	0.96	2.80	1.99	0.95
3	0.89	4.85	3.19	0.86	0.95	3.55	2.34	0.92	0.97	2.67	1.77	0.96
4	0.91	4.30	2.77	0.89	0.94	3.56	2.32	0.93	0.97	2.77	1.90	0.95
5	0.89	4.89	3.06	0.87	0.95	3.37	2.19	0.93	0.96	3.06	2.13	0.94
6	0.93	3.98	2.60	0.91	0.97	2.80	1.86	0.95	0.98	2.28	1.48	0.97
7	0.94	3.64	2.68	0.90	0.97	2.77	1.74	0.96	0.98	2.39	1.38	0.97
8	0.93	3.77	2.68	0.90	0.96	3.15	2.03	0.94	0.97	2.74	1.78	0.96
9	0.94	3.50	2.52	0.91	0.96	2.87	1.87	0.95	0.97	2.71	1.69	0.96
10	0.94	3.74	2.61	0.91	0.97	2.79	1.85	0.95	0.98	2.29	1.46	0.97
11	0.96	2.90	2.17	0.93	0.97	2.45	1.63	0.96	0.98	2.34	1.59	0.96
12	0.91	4.32	2.91	0.88	0.97	2.58	1.69	0.96	0.98	2.29	1.47	0.97
13	0.73	7.44	4.10	0.77	0.86	5.52	3.73	0.81	0.95	3.16	2.42	0.92
14	0.82	6.30	3.74	0.81	0.91	4.38	3.03	0.87	0.95	3.21	2.23	0.93
15	0.89	4.87	3.03	0.87	0.95	3.59	2.35	0.92	0.97	2.76	2.01	0.94
16	0.92	4.22	2.70	0.90	0.96	3.14	2.05	0.92	0.98	2.44	1.73	0.96

در ابتدا برای بهبود نتایج مدل‌سازی با روش RT از رویکردهای پیش‌پردازش Bagging و تبدیل موجک استفاده شد. نتایج به دست آمده از مدل‌سازی کمی WQI در جدول ۴ آورده شد.

با توجه به جدول ۴، نتیجه گرفته شد که استفاده از رویکرد پیش‌پردازش Bagging و تبدیل موجک باعث بهبود نتایج مدل‌سازی شده است. با توجه به این‌که رویکرد پیش‌پردازش Bagging با الگوریتم پایه درخت تصادفی، ترکیبی از نتایج چندین درخت تصادفی است، بنابراین استفاده از این روش، باعث افزایش دقت مدل RT شده است. پس به‌طور کلی نتیجه گرفته شد که استفاده از روش تبدیل موجک و دسته‌بندی، باعث افزایش دقت و کاهش خطا می‌شود.

از بین ترکیب‌های ورودی مورد مطالعه، تمامی ترکیب‌ها دقت قابل قبولی را داشتند. در روش RT سناریو ۱۱ شامل پارامترهای TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K, CO_3 با R=0.96 بالاترین دقت را در برآورد مقدار WQI داشته است. با ترکیب روش RT با رویکرد پیش‌پردازش Bagging و تبدیل موجک، مقدار R به ترتیب برابر ۰/۹۷ و ۰/۹۸ شده است که نشان از افزایش دقت مدل دارد. بهترین سناریو با بالاترین دقت و کمترین خطا مربوط به سناریو ۱۰ مدل W- B- RT با پارامترهای TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K است. نتایج نشان می‌دهد که تأثیر pH در برآورد مقدار عددی شاخص WQI کمتر از سایر پارامترها در نظر گرفته می‌شود. سناریو ۱ با یک ورودی (TH) دقت بالاتری از سناریو ۱۳ با دو ورودی (TH, K) داشته است. یعنی تاثیر پارامتر TH بیشتر از K بوده است.

برای درک بهتر تفاوت بین نتایج سناریوهای مختلف، مقدار خطای RMSE تمام سناریوهای مورد مطالعه در هر سه روش، در شکل ۴ نشان داده شد.



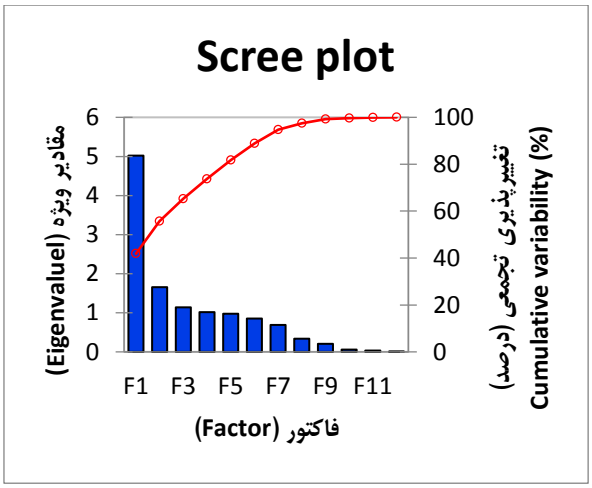
شکل ٤- مقادیر خطای RMSE روش‌های مورد مطالعه

Figure 4- RMSE error of the studied methods

همان‌طور که از جدول ۴ نیز مشاهده می‌شود؛ سناریو ۱۰ در روش W- B- RT کمترین مقدار RMSE را دارد که مقدار آن برابر ۲/۲۹ است. سناریوهای انتخاب شده با استفاده از روش همبستگی، نتایج بهتری از روش رلیف داشتند. می‌توان نتیجه گرفت که وجود همبستگی بین پارامترهای شیمیایی تاثیر به‌سزایی در مدل‌سازی عددی شاخص WQI دارد.

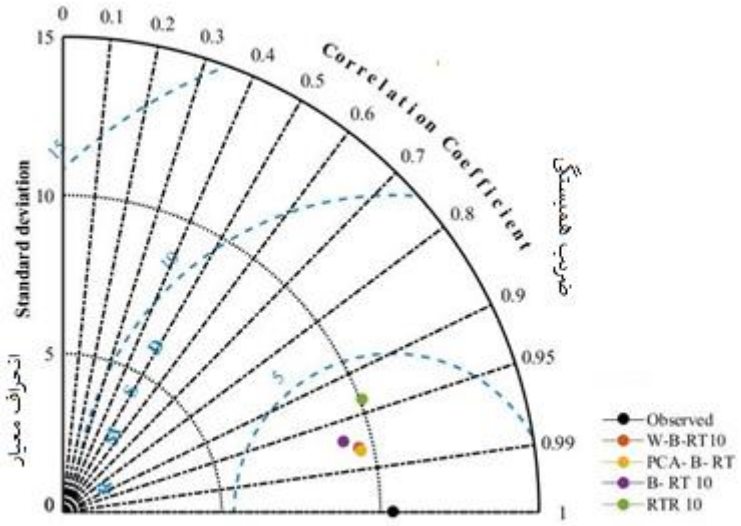
یکی دیگر از کارهایی که در پژوهش حاضر برای بهبود نتایج مدل‌سازی انجام شد، به کارگیری روش تحلیل مؤلفه اصلی بود. به کمک روش PCA، مؤثرترین ترکیبات خطی از پارامترهای ورودی شناسایی شده و به‌عنوان بردارهای ورودی، مورد استفاده قرار گرفتند. در نهایت عامل اصلی و تأثیرگذار استخراج شده از PCA، به‌عنوان ورودی روش Bagging با الگوریتم پایه RT در نظر گرفته شد. مقادیر ارزش ویژه و حالت تجمعی آن در نمودار Scree Plot نشان داده شد (شکل ۵). با توجه به این شکل نیز می‌توان نتیجه گرفت که از F1 به F12 با کاهش مقدار ارزش ویژه، تأثیر و ارزش عامل نیز کم شده است؛ بنابراین عامل F1، F2 و F3 به عنوان مؤلفه‌های اساسی انتخاب شدند.

با در نظر گرفتن ۳ عامل اصلی، مدل‌سازی انجام شد و مقدار R=0.98، RMSE=2.17، MAE=1.52 و Dr=0.97 به دست آمد.



شکل ۵- نمودار Scree Plot پارامترهای شیمیایی مورد مطالعه

Figure 5- Scree plot diagram of the studied chemical parameters



شکل ٦- دیاگرام تیلور برای بررسی تأثیر تبدیل موجک، رویکرد پیش‌پردازش Bagging و تحلیل مؤلفه اصلی

Figure 6- Taylor diagram to investigate the effect of wavelet transform, Bagging method and principal component analysis

همان طور که اشاره شد، سناریوی ۱۰ روش ترکیبی W-B- RT و روش PCA-B-RT بالاترین دقت و کمترین خطا را داشتند. برای نشان دادن بهتر تأثیر استفاده از رویکرد پیش‌پردازش Bagging و تبدیل موجک در سناریو ۱۰ و همچنین تأثیر روش تحلیل مؤلفه اصلی، در شکل ۶ دیاگرام تیلور روش‌های مذکور، ترسیم گردید.

با توجه به شکل ۶، ملاحظه گردید که استفاده از رویکرد پیش‌پردازش Bagging و تبدیل موجک، باعث بهبود نتایج درخت تصادفی شده است. همچنین در حالت کلی نتایج نشان داد که روش PCAعلی‌رغم کاهش بعد بردارهای ورودی و ساده‌سازی آن، می تواند دقت و سرعت عملکرد مدل را ارتقا بخشد و به‌عنوان بهترین روش برای تخمین مقدار عددی شاخص WQI معرفی شود.

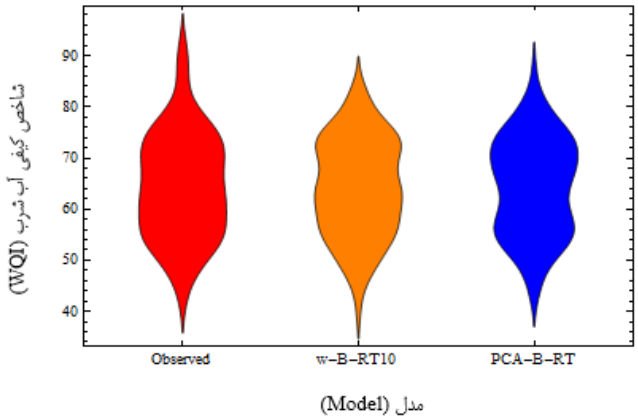
در شکل ۷ نیز نمودار ویلونی برای مقادیر شاخص WQI که به‌عنوان مقادیر مشاهداتی در نظر گرفته شدند؛ سناریوی برتر روش ترکیبی W-B- RT (سناریو ۱۰ شامل پارامترهای TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K) و روش PCA-B-RT رسم شد.

با توجه به شکل ۷، برآورد مقادیر حداقل و حداکثر شاخص WQI در مدل PCA-B-RT بهتر از مدل W-B-RT بوده است. همچنین بیشترین فراوانی WQI (بیشترین عرض نمودار ویلونی) در هر دو روش، حول چارک سوم تغییر نموده و نزدیک به مقادیر مشاهداتی برآورد شده است.

برای نشان دادن بهتر تغییرات داده‌های مشاهداتی و مدل برتر (PCA-B-RT) در شکل ۸ نمودار سری زمانی ارائه شد.

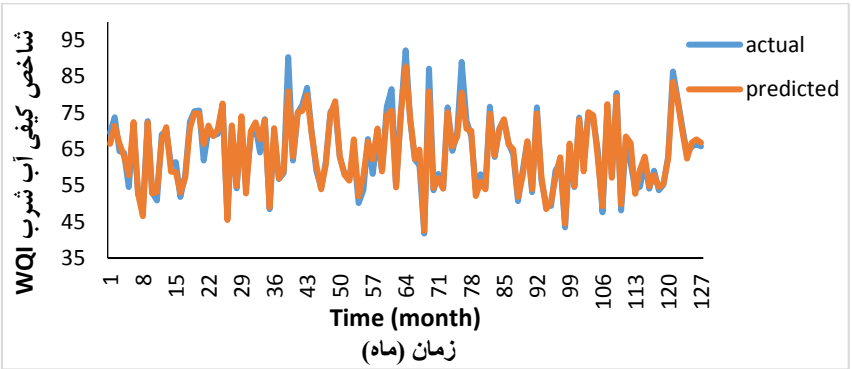
همان طور که از شکل ۸ مشخص است، روش مذکور با دقت بالایی شاخص WQI را برآورد کرده است.

بررسی منابع نشان می‌دهد، از روش‌های داده‌محور در مدل‌سازی کیفی آب به کرات استفاده گردیده و اکثر محققین بر توانمندی این مدل‌ها تاکید نموده‌اند.



شکل ۷- نمودار ویلونی سناریو برتر روش ترکیبی W-B-RT و روش PCA-B-RT

Figure 7- Villon diagram of the best scenario of the combined W-B-RT method and PCA-B-RT method



شکل ۸- نمودار سری زمانی داده‌های مشاهداتی و مدل برتر

Figure 8- Time series plot of observational data and superior model

آل مختار و یاسین (Al- Mukhtar and Al- Yaseen, 2019)، ستاری و همکاران (Sattari *et al*., 2017)، ضمن انجام مطالعاتی مشابه با مطالعه حاضر در حالت کلی به این نتیجه رسیدند که با عنایت به عملکرد مطلوب مدل‌های داده‌محور در مدل‌سازی پارامترهای کیفی آب، استفاده از این روش‌ها برای موارد مشابه قابل توصیه است. همچنین سلگی و همکاران (Solgi *et al*., 2017) برای تحلیل کیفیت آب رودخانه کارون واقع در غرب ایران طی مطالعه‌ای به مدل‌سازی و پیش‌بینی اکسیژن مورد نیاز بیولوژیکی پرداختند. نتایج تحقیق آن‌ها نشان داد که مدل SVM با ضریب تبیین ۰/۸۴ و جذر میانگین مربعات خطای ۰/۰۳۳۸ میلی‌گرم بر لیتر عملکرد نسبتا مطلوبی ارائه می‌کند. ایشان پس از اعمال تبدیل موجک روی داده های ورودی مدل، باعث شدند تا ضریب تبیین افزایش و به مقدار ۰/۹۴ و جذر میانگین مربعات خطا کاهش و به مقدار ۰/۰۲۱۰ میلی گرم بر لیتر برسد. بنابراین آن‌ها نتیجه گرفتند که ترکیب ماشین بردار پشتیبان با تبدیل موجک باعث بهبود نتایج پیش‌بینی مقدار BOD در رودخانه کارون می‌گردد. کرباسی و دیندار (Karbasi and Dindar 2019) نیز برای پیش‌بینی هدایت الکتریکی و نسبت جذب سدیم در رودخانه زاینده‌رود مدل‌های شبکه عصبی مصنوعی MLP و GMD را به‌صورت تکی و همراه با تبدیل موجک به کار بردند. نتایج به دست آمده از تحقیق آن‌ها نیز بیانگر بهبود عملکرد مدل‌ها در اثر استفاده از تبدیل موجک بود. همچنان که ملاحظه می‌گردد، با مقایسه نتایج تحقیق حاضر با سایر تحقیقات انجام یافته می‌توان دریافت که استفاده از تبدیل موجک و روش‌های داده‌مبنا نتایج قابل قبولی در بررسی کیفیت آب‌های سطحی داشته است. تفاوت تحقیق حاضر با سایر تحقیق‌های انجام شده در انتخاب نوع پارامتر هدف مورد مطالعه برای بررسی کیفیت آب بوده است. در تحقیقات بررسی شده فوق پارامترهایی از جمله اکسیژن مورد نیاز بیولوژیکی، هدایت الکتریکی و نسبت جذب سدیم با استفاده از روش‌های داده مبنا مدل‌سازی شده و تاثیر تبدیل موجک در بهبود نتایج مورد بررسی قرار گرفته است؛ اما در این پژوهش ابتدا شاخص کیفی آب با استفاده از ۱۲ پارامتر شیمیایی محاسبه شده و سپس با روش‌های داده‌مبنا مدل‌سازی شده و همانند تحقیقات مشابه فوق‌الذکر تاثیر تبدیل موجک در بهبود نتایج مورد ارزیابی و تاکید قرار گرفته است.

نتیجه‌گیری

در پژوهش حاضر، کیفیت آب ایستگاه هیدرومتری باغ کلایه مورد بررسی قرار گرفت. برای طبقه‌بندی کمی آب از ۱۲ پارامتر شیمیایی شامل pH، EC، TDS، Ca، Na، Mg، K، Cl، CO_3، HCO_3 و SO_4 در دوره آماری ۲۳ ساله در سناریوهای مختلف استفاده شد. ابتدا مقادیر کمی شاخص WQI محاسبه شد. پس از بررسی روند مقادیر عددی شاخص WQI، مشخص شد که این شاخص، روندی افزایشی داشته و در گذر زمان با افزایش مقدار عددی آن، کیفیت آب شرب ایستگاه مورد مطالعه کمتر شده است. با توجه به روند افزایشی هدایت الکتریکی در ایستگاه مورد مطالعه، نتیجه گرفته شد که افزایش شوری آب، می‌تواند به‌عنوان یکی از دلایل کاهش کیفیت آب در نظر گرفته شود. برآورد مقادیر عددی شاخص WQI با استفاده از روش درخت تصادفی انجام شد؛ همچنین تأثیر استفاده از رویکرد پیش‌پردازش Bagging، تبدیل موجک و تحلیل مؤلفه اصلی در بهبود نتایج، بررسی شد. ارزیابی روش‌ها با آماره‌های ضریب همبستگی، ریشه میانگین مربعات خطا، میانگین خطای مطلق و ضریب ویلموت اصلاح شده انجام گرفت. نتایج به دست آمده از پژوهش حاضر، نشان داد که استفاده از رویکرد پیش‌پردارش Bagging، تبدیل موجک و PCA در بهبود نتایج و افزایش دقت، تأثیر مثبتی داشتند. در برآورد مقادیر عددی شاخص WQI، روش PCA- B- RT با در نظر گرفتن ۳ عامل اصلی، بالاترین دقت را داشت. با توجه به این‌که تمام روش‌های مورد استفاده در برآورد مقادیر کمی ، دقت قابل قبولی داشتند، لذا در صورت کمبود داده و عدم دسترسی به تمام پارامترهای شیمیایی، می‌توان با استفاده از تعداد محدودی از پارامترها و روش‌های داده‌کاوی، نتایج مناسب و قابل قبولی را به دست آورد. با توجه به محدودیت تهیه داده و محدودیت صفحات، این مطالعه روی یک ایستگاه انجام گرفته است. جهت ارزیابی‌های دقیق‌تر، مطالعه روی ایستگاه متفاوت با طول دوره آماری زیاد منجر به نتایج جامع‌تر و کامل‌تری خواهد شد. این مطالعه در اقلیم خشک و نیمه‌خشک انجام گرفت در حالی‌که مطالعه روی اقلیم‌های متفاوت باعث ارائه نتایج تکمیلی می‌گردد. امکان استفاده از همه مدل‌ها از جمله مدل‌های یادگیری عمیق، برنامه‌ریزی بیان ژن و مقایسه آن‌ها وجود نداشت. لذا پیشنهاد می‌شود در مطالعات بعدی، این روش‌ها نیز مورد توجه قرار گیرند.

منابع

1. Abdi, H., & Williams, L.J. (2010). Principal component analysis, Wiley interdisciplinary reviews: *Computational Statistics 2*(4): 433-459.
2. 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. https://doi.org/10.1016/j.matpr.2021.02.065.
3. 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*(24). https://doi.org/10.3390/hydrology6010024.
4. Batur, E., & Maktav, D. (2019). Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey. *IEEE Transactions on Geoscience and Remote Sensing 57*(5): 2983–2989. http://doi.org/10.1109/TGRS.2018.2879024.
5. Breiman, L. (1996). Bagging predictors. *Machine Learning* 24: 123–140.
6. Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., & 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* 115454. http://doi.org/10.1016/j.watres.2019.115454.
7. 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. JMLR: W and P. Vol.32. 9 pages.
8. Hameed, M., Shargi, S., Yaseen, Z., Afan, H., Hussain, A., & Elshafie A. (2017). Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in a tropical region, Malaysia. *Neural Computing and Applications* 28: 893-905. https://doi.org/10.1007/s00521-016-2404-7.
9. Haar, A. (1910). The theory of orthogonal function systems. *Mathematical Annals 69*(3): 331-371. http://doi.org/10.1007/BF01456326.
10. Hosseini, H., Shakeri, A., Rezaei, M., Dashti Barmaki, M., & Shahraki, M. (2019). Application of water quality index (WQI) and hydro-geochemistry for surface water quality assessment, Chahnimeh reservoirs in the Sistan and Baluchestan Province. *Iranian Journal of Health and Environment 11*(4): 575-586.
11. Karbasi, M., & Dindar, S. (2019). Comparison of wavelet-MLP and wavelet-GMDH models in forecasting EC and SAR at Zayandeh-Rood River. *Environmental Sciences 16*(4): 135-152. (In Persian with English abstract)
12. Kavita, D., & Jagdish, S. (2012). *Water resources management and water quality, case of Bhopa l‰,* International Conference on Chemical, Ecology and Environmental Sciences (ICEES'2012) 17-18march, Bangkok.
13. Khalil, B., Ouarda, T., & St-Hilaire, A. (2011). Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. *Journal of Hydrology* 405: 277–287.
14. Kheirabadi, Kh., Fayazi, J., Roshanfekr, H., & Abdollahi-Arpanahi, R. (2017). Evaluation of the effectiveness of bootstrap aggregating sampling technique in the accuracy of the genomic best linear unbiased prediction method. *Iranian Journal of Animal Science 48*(4): 573-584. http://doi.org/10.22059/ijas.2018.248547.653596.
15. 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. https://doi.org/10.3390/w14101552.
16. Kira, K., & Rendell, L.A. (1992). *The feature selection problem: traditional methods and a new algorithm*. AAAI-92 Proceedings of the tenth national conference on Artificial intelligence, Menlo Park, California. 129-134.
17. Kolli, K., & Seshadri, R. (2013). Ground water quality assessment using data mining techniques. *International Journal of Computer Applications 76*(15): 39-45.
18. Lau, K.M., & Weng, H.Y. (1995). Climate signal detection using wavelet transform, How to make time-series sing, *Bulletin of the American Meteorological Society* 76: 2391-2402.
19. Mat Nawi, N., Chen, G., Jensen, T., & Abdanan Mehdizadeh, S. (2013). Prediction and classification of sugarcane Brix based on skin scanning using visible and shortwave near infrared. *Biosystems Engineering 115*(2): 154–161.
20. Nihalani, S.M., & Meeruty, A. (2020). Water quality index evaluation for major rivers in Gujarat. *Environmental Science and Pollution Research* 28: 63523–63531. http://doi.org/10.1007/s11356-020-10509-5.
21. Othman, F., Alaaeldin, M., Seyam, M., Ahmed, A., Teo, F., Ming, Fai, Ch., Afan, H., Sherif, M., Sefelnasr, A., & Shafie, A. (2020). Efficient river water quality index prediction considering a minimal number of inputs variables. *Engineering Applications of Computational Fluid Mechanics 14*(1): 751-763. https://doi.org/10.1080/19942060.2020.1760942.
22. Sattari, M.T., Mirabbasi, R., & Abbasgholi, M. (2017). The use of data mining in predicting the quality of surface water (case study: the rivers of the northern slopes of Sahand). *Ecohydrology 4*(2): 407-419. (In Persian)
23. Singh, D.F. (1992). Studies on the water quality index of some major rivers of Pune, Maharashtra. *Proceedings Academy Environmental Biology* 1: 61–66.
24. 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. (In Persian with English abstract)
25. Solgi, A., Pourhaghi, A., Zarei, H., & Ansari, H. (2017). Modeling and forecast biological oxygen demand (BOD) using combination support vector machine with wavelet transform. *Journal of Water and Soil 31*(1): 86-100.
26. Trabelsi, F., & Hadj Ali, S. (2022). Exploring machine learning models in predicting irrigation groundwater quality indices for effective decision making in Medjerda River Basin, Tunisia. *Sustainability* 14.

https://doi.org/10.3390/su14042341.
27. Vishwanath, V., Mahesh Kumar, N., & Wakif, A. (2021). *Haar wavelet scrutinization of heat and mass transfer features during the convective boundary layer flow of a nanofluid moving over a nonlinearly stretching sheet*. Partial Differential Equations in Applied Mathematics 4, https://doi.org/10.1016/j.padiff.2021.100192.