Journal of Research in Environmental Health

Volume 9, Issue 3, Autumn 2023

Prediction of groundwater nitrate variations using AdaBoost approach

Mansour Baziar
Assistant Professor, Department of Environmental Health Engineering, Ferdows Faculty of Medical Sciences, Birjand University of Medical Sciences, Birjand, Iran. (Corresponding author):
baziar.ehe@gmail.com

Received: 2023/03/30
Accepted: 2023/11/01

Document Type: Research article

Doi:10.22038/jreh.2023.73330.1604

ABSTRACT

Background and Purpose: Nitrates have long been considered indicative of drinking water quality and a critical concern for human health. The evolution of advanced models for water quality management has spurred decision-makers to incorporate artificial intelligence technologies into water quality planning. This study aims to employ the AdaBoost model, one of the cutting-edge models in water quality management, to predict nitrate concentrations in groundwater using pH and EC (Electrical Conductivity) as input variables.

Materials and Methods: Initially, the study analyzed the Pearson correlation matrix and subsequently determined the input variables for multiple AdaBoost models with varying hyperparameters. A sensitivity and dependence analysis of the model's input variables was conducted to assess their impact on nitrate prediction.

Results: The results obtained from the AdaBoost model reveal R-squared (R^2) values of 0.915 for the training dataset and 0.924 for the test dataset. Additionally, the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) scores for the training dataset were recorded as 1.02, 1.01, 0.823, and 7.3%, respectively. For the test dataset, these metrics were observed in the order of 0.228, 0.477, 0.375, and 3.2%. The model's sensitivity analysis identified the pH variable as the most influential factor in nitrate prediction.

Conclusion: The model analysis demonstrates that the proposed method performs well in predicting nitrate concentrations. This approach holds significant potential for implementation as an intelligent system for forecasting water quality parameters.

Keywords: AdaBoost, groundwater, Nitrate, sensitivity analysis, water quality

▶ **Citation:** Baziar M. Prediction of groundwater nitrate variations using AdaBoost approach. *Iranian Journal of Research in Environmental Health*. Autumn 2023; 9(3): 279-289.

پیش بینی تغییرات غلظت نیترات آب زیر زمینی با رویکرد AdaBoost

منصور بازیار
استادیار مهندسی بهداشت محیط، دپارتمان مهندسی بهداشت محیط، دانشکده علوم پزشکی فردوس، دانشگاه علوم پزشکی بیرجند، بیرجند، ایران. (نویسنده مسئول) :
baziar.ehe@gmail.com

تاریخ دریافت: ۱۴۰۲/۰۱/۱۰
تاریخ پذیرش: ۱۴۰۲/۰۸/۱۰
نوع مقاله: پژوهشی

چکیده

زمینه و هدف: نیترات همواره به عنوان یک شاخص کیفیت آب آشامیدنی و یک موضوع اساسی در سلامت انسان مورد توجه بوده است. توسعه مدل‌های پیشرفته برای مدیریت کیفیت آب، تصمیم‌گیرندگان را تشویق کرده است که فناوری‌های هوش مصنوعی را در برنامه‌ریزی کیفیت آب لحاظ نمایند. این مطالعه، قصد دارد تا با استفاده از مدل‌های AdaBoost (تقویت تطبیقی) بعنوان یکی از مدل‌های نوظهور در حیطه مدیریت کیفیت آب به پیش بینی غلظت نیترات در آب زیرزمینی با استفاده از هدایت الکتریکی، pH بپردازد.

مواد و روش‌ها: در این مطالعه ابتدا تحلیل همبستگی پیرسون انجام شد سپس با تعیین متغیرهای ورودی مدل چندین مدل AdaBoost با هایپر پارامترهای مختلف ساخته شد. سپس تحلیل حساسیت و وابستگی متغیرهای ورودی مدل در پیش بینی نیترات ارزیابی شدند.

یافته‌ها: نتایج مدل AdaBoostنشان داد که مقادیر ضریب R^2 برای داده آموزش ۰/۹۱۵ و برای داده‌های تست ۰/۹۲۴ بودند. مقادیر MSE، RMSE، MAE، MAPE برای داده‌های آموزش به ترتیب ۱/۰۲، ۱/۰۱، ۰/۸۲۳ و ۷/۳ درصد بدست آمد. این معیارها برای داده‌های تست به ترتیب ۰/۲۲۸، ۰/۴۷۷، ۰/۳۷۵ و ۳/۲ درصد بودند. تحلیل حساسیت مدل، متغیر pH به عنوان مهمترین متغیر تاثیر گذار در پیش بینی نیترات معرفی کرد.

نتیجه‌گیری: تحلیل مدل نشان داد که روش پیشنهادی در پیش بینی غلظت نیترات عملکرد بالایی دارد. این روش پتانسیل ویژه برای پیاده‌سازی به عنوان یک سامانه هوشمند برای پیش‌بینی پارامترهای کیفیت آب را دارد.

کلید واژه‌ها: AdaBoost، آب زیرزمینی، آنالیز حساسیت، کیفیت آب، نیترات.

استناد: بازیار م. پیش بینی تغییرات غلظت نیترات آب زیر زمینی با رویکرد AdaBoost. *فصلنامه پژوهش در بهداشت محیط*. پاییز ۱۴۰۲؛۹(۳): ۲۷۹-۲۸۹.

مقدمه

فرایند توسعه در یک منطقه منجر به صنعتی شدن آن، افزایش شهرنشینی و انقلاب در کشاورزی می‌شود که حاصل آن تخلیه آلاینده‌های مختلف به محیط زیست است. دفع بی رویه مواد شیمیایی صنعتی به محیط زیست تهدید قابل توجهی به منابع طبیعی است. آب‌های زیرزمینی منبع گرانبهای آب برای فعالیت‌های صنعتی و کشاورزی در دوران مدرن توسعه بشری است (۱). با این وجود، در کشورهای در حال توسعه، آب‌های زیرزمینی منبع قابل توجهی از آب آشامیدنی برای جمعیت روستایی و شهری محسوب می‌شود. اما فعالیت‌های مستمر انسانی و انقلاب صنعتی در چند دهه گذشته، کیفیت و کمیت منابع آب زیرزمینی را به شدت تهدید کرده است (۲). نیتروژن از طریق منابع سطحی مختلف مانند کودهای شیمیایی، دفع فضولات حیوانی، شیرابه محل دفن پسماند، فاضلاب شهری و غیره وارد آب‌های زیرزمینی می‌شود (۳). وجود نیترات در آب‌های زیرزمینی بویژه در کشورهای در حال توسعه به عنوان یک مسئله اساسی برای کیفیت آب‌های زیرزمینی و سلامتی انسان همیشه مطرح شده است و در برخی مناطق، غلظت آن به طور قابل توجهی بالاتر از غلظت‌های استاندارد تدوین شده برای آب آشامیدنی گزارش شده است (۴). کودهای نیتروژن دار به عنوان منبع اصلی آلودگی نیترات در خاک و همچنین آب‌های زیرزمینی شناخته می‌شوند (۵). نیترات دارای حلالیت بالا در آب و پایداری کمی در خاک دارد. اگر گیاهان از نیترات استفاده نکنند این عامل به نیتریت و گاز نیتروژن تبدیل شده و به راحتی به لایه زیرین خاک و در نهایت به آب‌های زیرزمینی منتقل می‌شود . نوشیدن آب حاوی نیترات بالا ممکن است خطرات زیادی برای سلامتی انسان ایجاد کند. به طور کلی، نوزادان در برابر آلودگی نیترات آسیب پذیرتر هستند، اما کودکان و بزرگسالان نیز ممکن است به دلیل مصرف آب غنی از نیترات با ناهنجاری‌هایی مانند اختلال عملکرد تیروئید مواجه شوند (۶). در طول دو دهه اخیر، کاربرد تکنیک‌های هوش مصنوعی در بسیاری از زمینه‌ها بویژه در زمینه‌های پیش‌بینی هیدرولوژیکی افزایش یافته است. برای مدل‌سازی آب‌های زیرزمینی، بیشتر تکنیک‌های هوش مصنوعی در پیش‌بینی سطح آب زیرزمینی بوده که نتایج رضایت‌بخشی نیز حاصل شده است (۷). در مورد پیش‌بینی کیفیت آب، مطالعات متعددی نیز با استفاده از تکنیک‌های هوش مصنوعی انجام شده است از جمله می‌توان به مطالعه تیاشا و یاسین (۲۰۲۰) اشاره کرد (۸). لو و ما (۲۰۲۰) از الگوریتم تقویت گرادیان و جنگل تصادفی (RF)[1] برای پیش‌بینی شش شاخص کیفیت آب در رودخانه توالاتین استفاده کردند (۹). کاستریو و گارسیا (۲۰۲۰) از مدل‌های خطی و RF برای تخمین غلظت مواد مغذی در رودخانه تیمز استفاده کردند (۱۰). مایرز و همکاران (۲۰۱۷) از مدل‌های شبکه عصبی مصنوعی (ANN)[2]، ماشین بردار پشتیبانی (SVM)[3] و مدل‌های RF برای پیش‌بینی کدورت آب یک شبکه توزیع آب شاخه‌ای در بریتانیا استفاده کردند (۱۱). ال بلالی و همکاران (۲۰۲۰) از مدل‌های ANN برای پیش بینی کیفیت شیمیایی آب زیرزمینی برای اهداف آشامیدنی استفاده کرد (۱۲). فیجانی و همکاران (۲۰۱۹) روش‌های مختلف هوش مصنوعی را برای پشتیبانی و پایش برخط کیفیت آب مخزن طراحی و اجرا کردند (۱۳). واگ و همکاران (۲۰۱۶) از مدل ANN برای تعیین مناسب بودن کیفیت آب زیرزمینی برای اهداف آبیاری استفاده کردند آنها در این مطالعه از ۱۳ پارامتر فیزیکی-شیمیایی آب استفاده کردند و عملکرد عالی مدل توسعه داده شده را شاهد بودند (۱۴). نکته مهم همه این مطالعات این است که مدل‌های هوش مصنوعی در پیش‌بینی و ارزیابی کیفیت آب بسیار دقیق هستند. بهرحال، کارایی مدل هوش مصنوعی تنها به دقت پیش‌بینی بستگی ندارد، بلکه به ماهیت و تعداد متغیرهای مورد استفاده نیز بستگی دارد. در این راستا، افزایش متغیرهای ورودی باعث کاهش کارایی مدل‌های هوش مصنوعی شده و کاربرد میدانی آنها را نیز تضعیف می‌کند(۱۲). بررسی متون علمی نشان

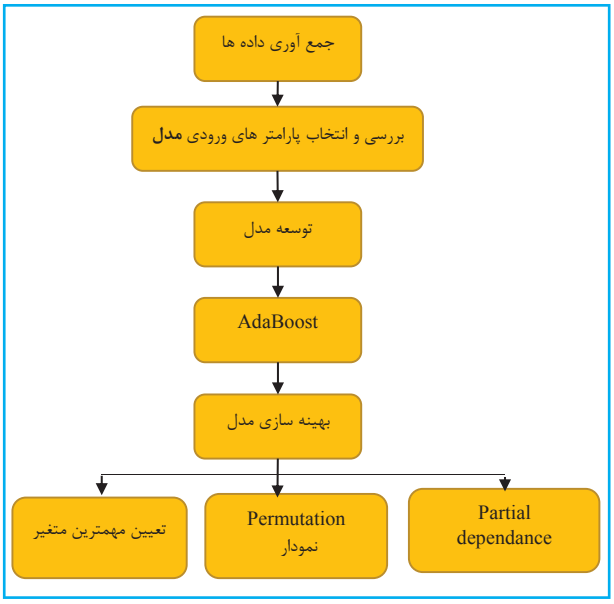
1. Random Forest
2. Artificial Neural Network
3. Support Vector Machine

می‌دهد که پارامترهای ورودی اکثر مدل‌های توسعه داده شده برای پیش بینی یک شاخص کیفیت آب زیاد بوده و همچنین این پارامترها باید به طور دقیق توسط کارشناس خبره در آزمایشگاه ارزیابی شوند که این عمل باعث افزایش هزینه‌های سنجش (هزینه مواد شیمیایی، انسانی) شده و حتی ممکن است منجر به خطای سنجش نیز شود (۷). علاوه بر این، قدرت تعمیم و حساسیت مدل‌های هوش مصنوعی به متغیرهای ورودی به اندازه کافی تجزیه و تحلیل نشده است. در این راستا استفاده از پارامترهای فیزیکی مثل هدایت الکتریکی، pH که می‌توان آنها را با فناوری‌های حسگر سنجید و از آنها بعنوان ورودی مدل استفاده کرد (۱۴)، می‌تواند علاوه بر کاهش هزینه‌ها و کاهش خطاهای انسانی، کارایی مدل‌های هوش مصنوعی را به میزان قابل توجهی بهبود بخشید. این عمل، مسئولان و تصمیم گیرندگان را تشویق می‌کند تا فناوری‌های هوش مصنوعی را برای برنامه ریزی و مدیریت کیفیت آب پیاده سازی نمایند (۱۴). بر این اساس به نظر می‌رسد که توسعه مدل‌های هوش مصنوعی برای پیش‌بینی شاخص‌های کیفیت آب آشامیدنی با استفاده از داده‌های آرشیو شده بویژه از پارامترهای فیزیکی حیاتی باشد. در نتیجه، این مطالعه، قصد دارد تا با استفاده از مدل‌های AdaBoost[1] (تقویت تطبیقی) بعنوان یکی از مدل‌های نوظهور در حیطه مدیریت کیفیت آب به پیش بینی غلظت نیترات در آب زیرزمینی با استفاده از هدایت الکتریکی، pH بپردازد. همچنین تحلیلی بر حساسیت مدل‌ها به پارامترهای ورودی ارائه خواهد شد.

روش کار

مدل AdaBoost

مدل AdaBoost یک الگوریتم تقویتی یادگیری ماشین است که برای مسائل طبقه بندی و بویژه برای مسائل رگرسیون استفاده می‌شود. این الگوریتم با استفاده از ترکیب چندین دسته‌بند ضعیف، یک دسته‌بند قوی تولید می‌کند. مفهوم ریاضی اصلی در AdaBoost، وزن‌دهی به نمونه‌ها است. در هر مرحله از الگوریتم، وزن هر نمونه بر اساس عملکرد دسته‌بند قبلی تعیین می‌شود. نمونه‌هایی که در مراحل قبل به درستی یا با دقت بیشتری تشخیص داده شده‌اند، وزن کمتری در مرحله جدید خواهند داشت. در عوض، نمونه‌هایی که دارای پاسخ اشتباه بودند، وزن بیشتری در مرحله جدید خواهند داشت. سپس، چندین دسته‌بند ضعیف با استفاده از این وزن‌دهی، برای دسته‌بندی داده‌ها آموزش داده می‌شوند. در هر مرحله از دسته‌بندی بهترین دسته‌بند ضعیف انتخاب می‌شود و وزن آن برای تولید دسته‌بند قوی بعدی استفاده می‌شود. در نهایت، با ترکیب تمامی دسته‌بندهای ضعیف با وزن‌دهی مناسب، یک دسته‌بند قوی تولید می‌شود که برای دسته‌بندی داده‌های جدید استفاده می‌شود (۷). در این مطالعه از نرم افزار پایتون برای مدل سازی استفاده شده است. فرایند انجام مطالعه در شکل ۱ نشان داده شده است.



شکل ۱. فرایند توسعه مدل AdaBoost و تحلیل آن

جمع آوری و پردازش داده

داده‌های این مطالعه از بررسی مقالات در جورنال‌های مختلف و از طریق موتور جستجو گوگل اسکالر حاصل شد. بدین منظور تیم تحقیقاتی در جستجوی یافتن مقاله یا مقاله‌هایی بودند که به‌طور

1. Adaptive Boosting

جامع اطلاعات شیمیایی آب را منتشر کرده باشد، در نهایت از داده‌های این مقاله برای مدل سازی استفاده گردید (۱۵). در مقاله یافت شده داده‌های کیفیت آب زیر زمینی مربوط به ۳۳ چاه در کشور عراق گزارش شده بود. پارامترهای کیفی گزارش شده شامل pH، کل جامدات محلول، EC، کلسیم، منیزیم، کلراید، نیترات، پتاسیم، سولفات، بی کربنات و نسبت جذب سدیم[1] بودند (۱۵). با توجه به اینکه هدف توسعه یک مدل مناسب برای پیش بینی نیترات بود، یک بررسی عمیق در انتخاب ورودی مدل انجام گردید. در انتخاب متغیرهای ورودی مدل، معیارهای سهولت اندازه گیری، کاهش خطای انسانی در آزمایش و حداقل هزینه انجام آزمایش لحاظ شده بودند. علاوه براین یک تحلیل همبستگی پیرسون نیز برای شناسایی همبستگی متغیرها با نیترات و دیگر پارامترها انجام شد. بعد از شناسایی این پارامترها، داده‌های خام در مقادیر ۰/۱۰ تا ۰/۹ نرمال گردید (معادله ۱). دلیل نرمال سازی کاهش پیچیدگی محاسبات برای الگوریتم مد نظر بود. همچنین نرمال سازی باعث می‌شود که وزن‌های اعمالی به متغیر، همسان سازی شود تا تحلیل مهمترین متغیر تاثیرگذار در مدل به شکل صحیحی انجام شود. علاوه براین به جهت ارزیابی قدرت مدل ها، داده‌ها به دو دسته آموزش و تست با سهم به ترتیب ۸۰ درصد و بیست درصد تقسیم شدند (۱۶).

معادله ۱

$$y = \frac{x_i - x_{min}}{x_{max} - x_{min}} \times (b - a) + a$$

در معادله فوق، a و b به ترتیب مقادیر ۰/۱ و ۰/۹ هستند و x داده‌های مدل هستند. اندیس‌های min و max به حداقل و حداکثر مقدار داده‌های هر متغیر اشاره دارد.

معیارهای ارزیابی مدل

برای ارزیابی مدل‌های توسعه داده معیارهای مختلفی از جمله میانگین مربعات خطا (MSE)[2]، ریشه میانگین مربعات خطا (RMSE)[3]، میانگین خطای مطلق (MAE)[4]، میانگین درصد خطای مطلق (MAPE)[5] و ضریب تبیین (R^2) استفاده شد. در ارزیابی معیارها مدلی بهینه خواهد بود که حداقل مقادیر خطاهای ذکر شده فوق و حداکثر مقدار ضریب تبیین را داشته باشد. بطور معمول ضریب تبیین نسبت تغییرات متغیر وابسته را که می‌توان به متغیر مستقل نسبت داد، اندازه گیری می‌کند. معیارهای ارزیابی مدل‌های توسعه داده شده بر اساس معادلات ۲ تا ۶ انجام شده است. در این معادلات act_i و pre_i به مقدار واقعی نیترات و مقدار پیش بینی آن اشاره دارند. همچنین $\overline{act}$ و $\overline{pre}$ میانگین داده‌های واقعی و پیش بینی را بیان می‌کنند.

معادله ۲

$$MAE = \frac{\sum_{i=1}^{N} |act_i - pre_i|}{N}$$

معادله ۳

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n} (act_i - pre_i)^2}$$

معادله ۴

$$MSE = \frac{1}{N}\sum_{i=1}^{n} (act_i - pre_i)^2$$

معادله ۵

$$R^2 = \left(\frac{\sum_{i=1}^{n} (act_i - \overline{act})(pre_i - \overline{pre})}{\sqrt{\sum_{i=1}^{n} (act_i - \overline{act})^2 \sum_{i=1}^{n} (pre_i - \overline{pre})^2}} \right)^2$$

معادله ۶

$$MAPE = \frac{1}{N}\sum_{i=1}^{n} \left| \frac{act_i - pre_i}{act_i} \right| \times 100$$

یافته ها

نتایج آماری و تحلیل ماتریس همبستگی پیرسون

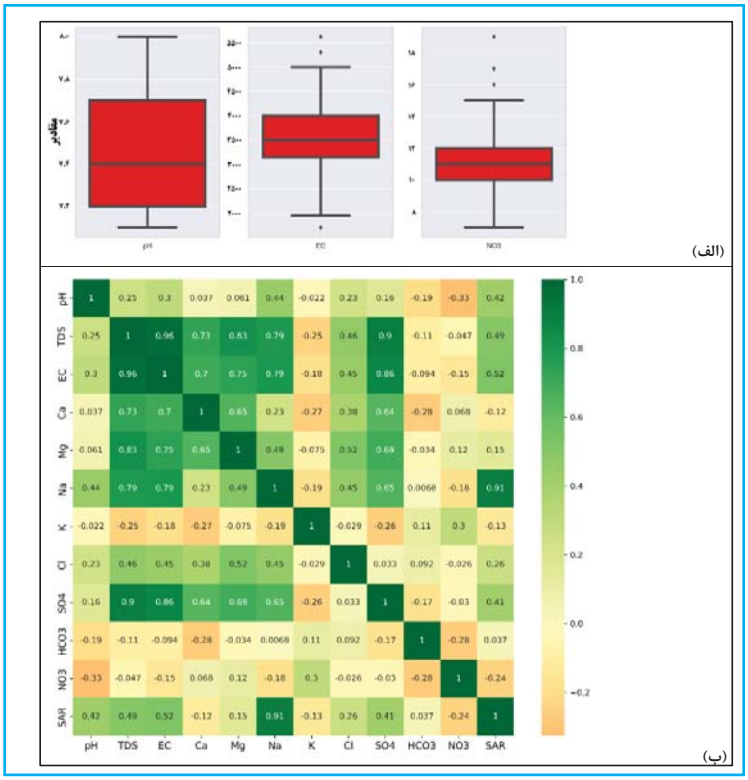
تحلیل توصیفی کل پارامترها و همچنین توزیع آماری پارامترهای مورد استفاده در این مطالعه به ترتیب در جدول ۱ و شکل ۲ (الف) ارائه شده‌است. در آمار توصیفی، محدوده بین چارکی (IQR)[6] معیاری برای سنجش پراکندگی آماری داده‌ها است. نمودار جعبه‌ای یک روش

1. Sodium Absorption Ratio
2. Mean Squared Error
3. Root Mean Squared Error
4. Mean Absolute Error
5. Mean Absolute Percentage Eerror
6. Interquartile Range

استاندارد برای نمایش توزیع داده‌ها است که براساس شاخص‌های آماری «کوچکترین مقدار» ، «چارک اول»، «میانه، «چارک سوم» و «بزرگترین مقدار» ساخته شده است (پنج خط از پایین به بالای شکل ۲ (الف) این شاخص‌ها را بیان می‌کند). همچنین این نمودار می‌تواند در مورد وجود داده‌های پرت، اطلاعاتی را ارائه و مقادیر آن‌ها را تعیین نماید (۱۷). شکل ۲(ب) نتایج تحلیل ماتریس همبستگی پیرسون متغیرها را نشان می‌دهد. در این شکل روابط بین متغیرها با اعداد بین ۱- تا ۱ نشان داده شده است. اگر رابطه بین دو متغیر دارای عدد ۱- باشد نشان دهنده همبستگی خطی کاملاً منفی بین دو متغیر است. عدد صفر نمایانگر عدم همبستگی خطی بین دو متغیر است و عدد یک نشان دهنده همبستگی خطی کاملا مثبت بین دو متغیر است. هر چه ضریب پیرسون از مقدار صفر دورتر شود، رابطه بین دو متغیر قوی تر می‌شود (۱۷).

جدول ۱. نتایج آنالیز توصیفی تمام متغیرها (تمام واحدها برحسب ppm بجز EC برحسب µS/cm)

آنالیز ها	SAR	NO_3	HCO_3	SO_4	Cl	K	Na	Mg	Ca	EC	TDS	pH
میانگین	۳/۵۴	۱۱/۳۹	۲۱/۴۴	۱۲۷۶/۸۸	۲۵۶/۳۶	۵/۴۲	۲۷۹/۱۸	۱۱۶/۷۹	۲۸۷/۸۵	۳۶۰۹/۰۹	۲۲۶۷/۶۷	۷/۴۵
خطای استاندارد	۰/۲۳	۰/۴۵	۱/۶۷	۶۳/۸۴	۲۷/۰۶	۰/۴۱	۱۸/۶۷	۷/۳۳	۱۳/۴۴	۱۵۶/۵۴	۱۰۱/۱۲	۰/۰۵
میانه	۳/۵۷	۱۱/۰۰	۲۱/۰۰	۱۲۶۷/۰۰	۲۰۶/۰۰	۵/۰۰	۲۹۸/۰۰	۱۰۸/۰۰	۲۹۷/۰۰	۳۵۰۰/۰۰	۲۲۲۰/۰۰	۷/۴۰
انحراف معیار	۱/۳۱	۲/۵۹	۹/۵۸	۳۶۶/۷۳	۱۵۵/۴۳	۲/۳۷	۱۰۷/۲۳	۴۲/۱۰	۷۷/۲۳	۸۹۹/۲۸	۵۸۰/۹۱	۰/۲۸
حداقل	۱/۴۰	۷/۰۰	۷/۸۰	۵۷۶/۰۰	۶۳/۰۰	۲/۰۰	۱۱۵/۰۰	۴۳/۰۰	۱۴۵/۰۰	۱۷۰۰/۰۰	۱۰۵۰/۰۰	۷/۱۰
حداکثر	۶/۴۳	۱۹/۰۰	۴۷/۰۰	۲۳۱۳/۰۰	۶۳۴/۰۰	۱۱/۰۰	۵۰۰/۰۰	۲۰۹/۰۰	۴۳۳/۰۰	۵۶۲۰/۰۰	۳۶۸۰/۰۰	۸/۰۰
تعداد	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰	۳۳/۰۰



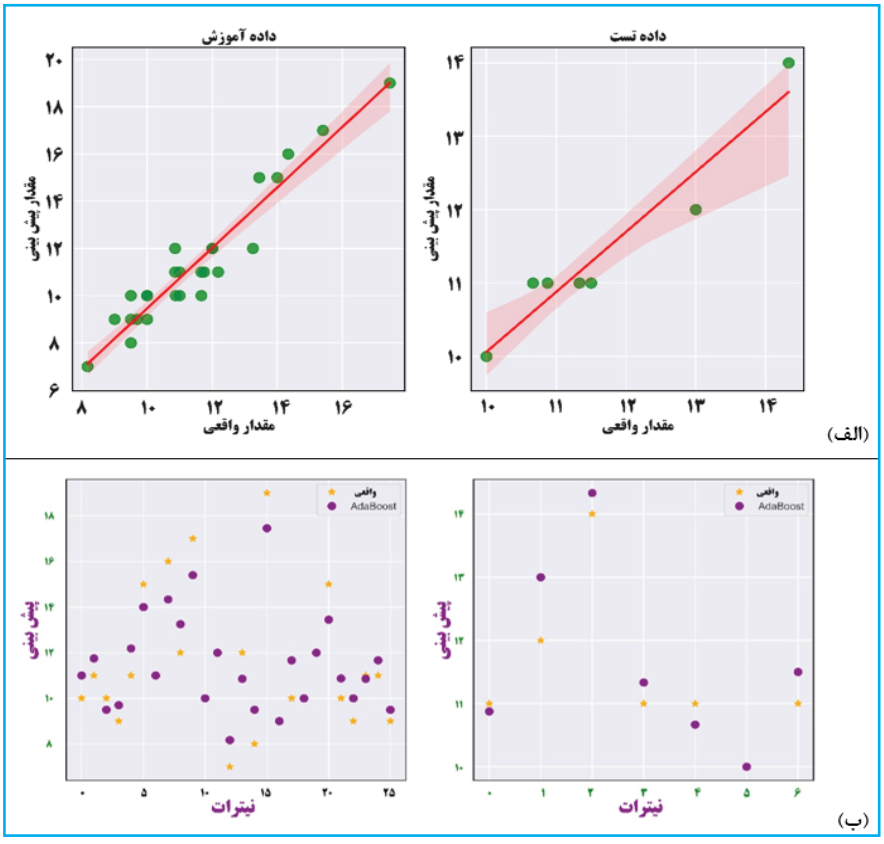
شکل ۲. نمودار جعبه‌ای متغیرهای مورد مطالعه در این تحقیق (الف) ماتریس همبستگی پیرسون بین متغیرها (ب)

بهینه سازی پارامترهای تنظیمی مدل AdaBoost

در این مطالعه به منظور بهینه سازی هایپر پارامترهای مدل AdaBoost موارد: نرخ یادگیری (۰/۰۰۱ تا ۱)، تعداد برآوردکننده‌ها (۵۰ تا ۳۰۰) و انواع تابع خطا (خطی، درجه ۲ و نمایی) مورد آزمون قرار گرفتند. در نهایت برای مدل بهینه نرخ یادگیری، تعداد برآوردکننده‌ها و تابع خطای به ترتیب ۰/۶، ۸۵ و تابع نمایی حاصل شد.

بررسی دقت و اعتبارسنجی مدل AdaBoost

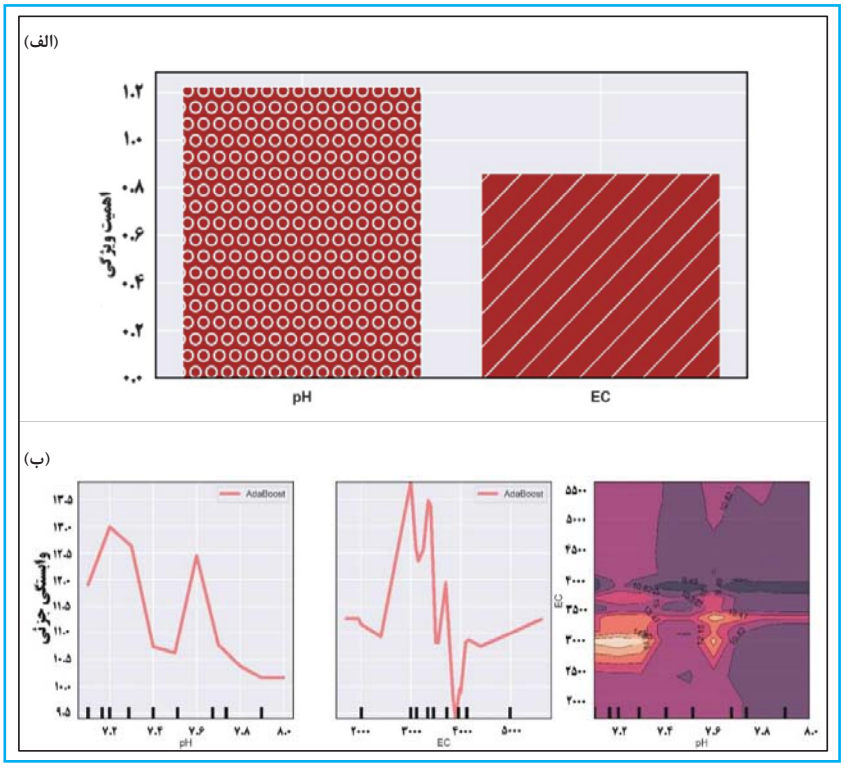
همانطور که پیش تر بیان شد؛ در مجموع بیست درصد داده‌ها برای فرآیند اعتبارسنجی مدل و هشتاد درصد داده‌ها برای آموزش مدل AdaBoost استفاده شده است. عملکرد مدل‌های توسعه داده با معیارهای مختلف (معادلات ۲-۶) مورد آزمون قرار گرفتند. شکل ۳ (الف) نمودارهای پراکندگی مقادیر پیش بینی شده غلظت نیترات در برابر مقادیر واقعی آن را نشان می‌دهد. ضریب تبیین بعنوان شاخصی برای همبستگی داده های مقادیر پیش بینی شده و داده های واقعی استفاده شد. با توجه به نمودار ۳ (الف) ضریب R^2 برای داده آموزش ۰/۹۱۵ و برای داده‌های تست ۰/۹۲۴ بدست آمد. بنابراین می‌توان نتیجه گرفت که مدل توسعه یافته قادر است، مقادیر غلظت نیترات را با دقت بالای ۹۱ درصد برای داده آموزش و با دقت بالای ۹۲ درصد برای داده‌های تست را پیش بینی نماید. همچنین در مدل بهینه مقادیر میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE)، میانگین خطای مطلق (MAE)، میانگین درصد خطای مطلق (MAPE) برای داده‌های آموزش به ترتیب ۱/۰۲، ۱/۰۱، ۰/۸۲۳ و ۷/۳ درصد بدست آمد. این معیارها برای داده‌های تست به ترتیب برابر با ۰/۲۲۸، ۰/۴۷۷، ۰/۳۷۵ و ۳/۲ درصد بود. نمودار ۳ (ب) نیز نزدیکی مقادیر پیش بینی شده غلظت نیترات توسط مدل AdaBoost و مقادیر واقعی آن را توصیف و تصدیق می کند.



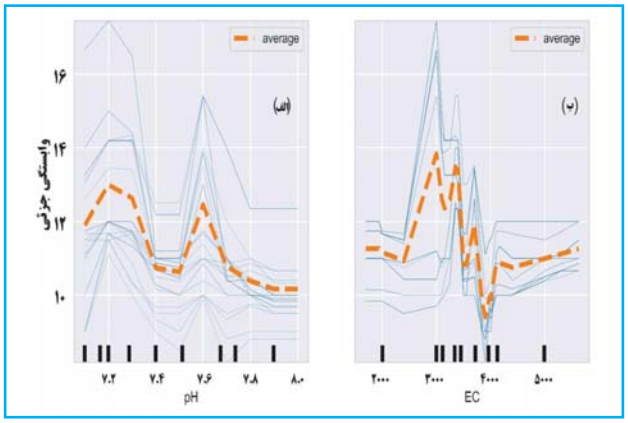
نمودار ۳. نمودارهای پراکندگی مقادیر پیش بینی شده غلظت نیترات در برابر مقادیر واقعی (الف). مقادیر واقعی در برابر پیش بینی (ب)

شناسایی مهمترین متغیر تاثیرگذار در مدل AdaBoost و تحلیل نمودار Partial dependance

اهمیت نسبی هر متغیر بر اساس روش Permutation و تحلیل نمودار Partial dependance پس از توسعه مدل و شناسایی مدل بهینه انجام گردید. درحقیقت تحلیل اهمیت نسبی هر متغیر کمک می‌کند تا سهم و تاثیر پارامترهای pH و هدایت الکتریکی در پیش بینی غلظت نیترات شناسایی شوند (۱۸). شکل ۴ (الف) گویای تاثیر بیشتر پارامتر pH است. از طرفی نمودار Partial dependance (شکل ۴ (ب)) وابستگی این پارامترها به غلظت نیترات را نشان می‌دهد. شکل ۵ که به نمودار انتظار شرطی فردی [1] معروف است، وابستگی تک‌تک نمونه‌ها و میانگین وابستگی آنها به نیترات را نشان می‌دهد.



شکل ۴. نمودار تحلیل حساسیت متغیرهای هدایت الکتریکی و pH در پیش بینی غلظت نیترات (الف). نمودار Partial dependance (ب)



شکل ۵ . نمودار ICE نیترات بر روی پارامترهای pH و هدایت الکتریکی

1. Individual Conditional Expectation

بحث

آنالیز توصیفی اطلاعات مفیدی پیرامون وضعیت داده‌ها ارائه می‌دهد. همانطور که در جدول ۱ نشان داده شده است رنج تغییرات داده‌ها برای سه متغیر مورد مطالعه بسیار وسیع است. این شرایط ایجاب می‌کند که برای فرایند مدل سازی باید داده‌ها نرمال سازی شوند تا تحلیل‌های مناسب و صحیحی از مدل سازی فراهم شود. از طرفی آنالیز توصیفی کمک می‌کند تا از وجود داده‌های فراتر از حد استاندارد بویژه برای نیترات که اثرات بهداشتی دارد، شناسایی شوند. تحلیل ماتریس همبستگی پیرسون برای شناسای و درک ارتباط خطی بین متغیرها و نیترات استفاده شد (۱۹). از این تحلیل برای انتخاب متغیرهای مدل مد نظر استفاده گردید. همانطور که در شکل تحلیل ماتریس همبستگی نشان داده شده است و با توجه به معیارهای در نظر گرفته شده در این مطالعه سه پارامتر pH و هدایت الکتریکی و کل جامدات محلول در مرحله اول شناسایی شدند. اما وجود همبستگی بالا بین متغیرهای ورودی مدل برای یک مدل سازی صحیح مناسب نمی باشد. لذا با توجه به اینکه دو پارامتر هدایت الکتریکی و کل جامدات محلول دارای ضریب همبستگی ۰/۹۶ بودند (همبستگی بسیار بالا)، باید یکی از این پارامترها از مطالعه خارج می‌شد. لذا برای این کار از شکل تحلیل ماتریس همبستگی رابطه بین این پارامترها با نیترات ارزیابی شد. متغیری بر طبق متون علمی انتخاب می‌شود که دارای همبستگی بیشتری با نیترات داشته باشد. لذا پارامتر هدایت الکتریکی به دلیل داشتن ضریب همبستگی بیشتر انتخاب و پارامتر کل جامدات محلول از مطالعه خارج شد. AdaBoost یک الگوریتم یادگیری ماشینی است که برای بهبود دقت مدل از چندین مدل ضعیف استفاده می‌کند. هر مدل ضعیف در AdaBoost به عنوان یک "توابع تصمیم‌گیری ساده" تعریف می‌شود. هایپر پارامترهای AdaBoost شامل تعداد مدل‌های ضعیف، نرخ یادگیری و تعداد ایپاک‌ها است. بهینه سازی هایپر پارامترهای AdaBoost بسیار مهم است و می‌تواند بهبود قابل توجهی در دقت مدل داشته باشد. در بهینه سازی تعداد مدل‌های ضعیف، باید به دقت و سرعت یادگیری توجه کرد (۲۰). اگر تعداد مدل‌های ضعیف زیاد باشد، ممکن است به دقت مدل کمک کند، اما سرعت یادگیری را کاهش می‌دهد. نرخ یادگیری نیز بسیار مهم است. اگر نرخ یادگیری بسیار بالا باشد، ممکن است مدل به سرعت به حالت بیش‌برازش برسد و در داده‌های جدید دقت پایینی داشته باشد. از طرف دیگر، اگر نرخ یادگیری کم باشد، ممکن است به سرعت به حالت زیربهینه برسد و دقت مدل را کاهش دهد (۲۰). برای پیدا کردن بهترین نرخ یادگیری، نیاز است که از روش‌های اعتبارسنجی و آزمون‌های مختلف استفاده شود. در کل، بهینه سازی هایپر پارامترهای AdaBoost نیازمند تجربه و تخصص در زمینه یادگیری ماشینی است. باید به دقت به هر پارامتر توجه شود و از روش‌های اعتبارسنجی مناسب برای پیدا کردن بهترین مقادیر آن‌ها استفاده شود. در این مطالعه هایپر پارامترهای نرخ یادگیری، تعداد برآوردکننده‌ها و تابع خطای به ترتیب ۰/۶، ۸۵ و تابع نمایی حاصل شد و ارزیابی عملکرد مدل‌ها نشان داد که مدل توسعه داده شده از دقت بالایی (شکل ۳) در برآورد غلظت نیترات دارد. علاوه بر این میانگین درصد خطای مطلق (MAPE) برای داده‌های آموزش ۷/۳ درصد و برای داده‌های تست ۳/۲ درصد بود که نشان می‌دهد خطای کلی مدل بسیار کم است. نتایج این تحقیق مطابقت بالایی با نتایج محققانی از جمله تیاشا و یاسین (۲۰۲۰)، لو و ما (۲۰۲۰)، کاستریو و گارسیا (۲۰۲۰)، مایرز و همکاران (۲۰۱۷) ال بلالی و همکاران (۲۰۲۰) فیجانی و همکاران (۲۰۱۹ واگ و همکاران (۲۰۱۶) داشت (۷-۱۱, ۱۳). عوامل فیزیکو- شیمیایی مختلفی بر غلظت نیترات و پایداری و تبدیل آن به اشکال دیگر در آب‌های زیرزمینی تأثیر می‌گذارند که pH و هدایت الکتریکی از این موارد هستند. در pHهای قلیایی یون‌های نیترات پایدارتر هستند و احتمال کمتری دارد که به اشکال گازی تبدیل شود. شرایط اسیدی می‌تواند منجر به تبدیل نیترات به گاز نیتروژن و کاهش غلظت آن در آب شود (۲۱). هدایت الکتریکی؛ اندازه گیری توانایی آب در هدایت الکتریسیته است که تحت تأثیر غلظت یون‌ها در آب

است. سطوح بالای آن می‌تواند نشان دهنده غلظت بالای یون‌ها باشد که می‌تواند جذب نیترات گیاه را کاهش و باعث افزایش ورود آن به آبهای زیر زمینی شود (۲۲). نتایج آنالیز حساسیت مدل بر طبق روش Permutation (شکل ۴ الف) نشان داد که هر دو پارامتر تاثیر قابل توجهی در پیش بینی غلظت نیترات دارند. بهرحال اهمیت نسبی پارامتر pH بیشتر از هدایت الکتریکی بود. نمودار Partial dependence (شکل ۴ ب) وابستگی متغیر pH و هدایت الکتریکی و برهمکنش آنها را در غلظت‌های مشاهده شده در این مطالعه را نشان می‌دهد. همانطور که از این شکل مشهود است بیشترین غلظت نیترات در مقادیر pH و هدایت الکتریکی برابر با ۷/۲ و ۳۰۰۰ µS/cm مشاهده می‌شود. نمودار ICE (شکل ۵) نموداری مشابه با Partial dependence است البته با این تفاوت که علاوه بر اینکه تاثیر میانگین یک متغیر را نشان می‌دهد قادر است اطلاعات اضافی در مورد تک تک نمونه‌ها فراهم کند. از جمله کاهش شدید وابستگی نیترات به pH که در pHهای بین ۷/۴ تا ۷/۵ مشاهده می‌شود و این درحالی است که در نمودار Partial dependence این خط صاف تر است. همچنین در نمودار مربوط به هایت الکتریکی نمونه‌ای وجود دارد که وابستگی آن در مقادیر ۴۰۰۰ تا ۶۰۰۰ واحد تغییری نمی کند.

نتیجه گیری

این مطالعه، با استفاده از مدل‌های AdaBoost به عنوان یکی از مدل‌های نوظهور در مدیریت کیفیت آب به پیش‌بینی غلظت نیترات در آب زیرزمینی پرداخت. مطالعه ابتدا با تحلیل همبستگی پیرسون شروع و سپس متغیرهای ورودی برای ایجاد مدل‌های AdaBoost با پارامترهای مختلف مشخص شدند. نتایج نشان دادند که مدل AdaBoost با ضرایب R^2 به ترتیب ۰/۹۱۵ برای داده‌های آموزش و ۰/۹۲۴ برای داده‌های تست، عملکرد بسیار قابل قبولی از خود نشان دادند. مقادیرپایین MSE، RMSE، MAE و MAPE نیز عملکرد مؤثر مدل را تایید کردند. این مقادیر برای داده‌های آموزش به ترتیب ۱/۰۲، ۱/۰۱، ۰/۸۲۳ و ۷/۳ درصد و برای داده‌های تست به ترتیب ۰/۲۲۸، ۰/۴۷۷، ۰/۳۷۵ و ۳/۲ درصد بودند. تحلیل حساسیت مدل، فرآیندی است که در آن به ارزیابی وابستگی مدل توسعه داده شده به تغییرات مقادیر ورودی آن اشاره دارد و به تعیین اهمیت نسبی هر متغیر ورودی در پیش‌بینی یا توصیف خروجی مدل کمک می‌کند، نشان داد که متغیر pH به عنوان مهمترین متغیر در پیش‌بینی غلظت نیترات انتخاب شده است. با توجه به نتایج مطلوب این مدل، می‌توان نتیجه گرفت که روش پیشنهادی برای پیش‌بینی غلظت نیترات عملکرد مناسبی داشته و دارای پتانسیل اجرایی به عنوان یک سامانه هوشمند در زمینه پیش‌بینی پارامترهای کیفیت آب را دارا است.

ملاحظات اخلاقی

نویسنده تمام نکات اخلاقی شامل سرقت ادبی، انتشار دوگانه، تحریف داده‌ها وداده سازی را در این مقاله رعایت کرده است. همچنین هرگونه تضاد منافع حقیقی یا مادی که ممکن است بر نتایج یا تفسیرمقاله تأثیر بگذارد را رد می کند.

تقدیر و تشکر

مطالعه حاضر با کد اخلاق IR.BUMS.REC.1402.193 در معاونت پژوهشی دانشگاه علوم پزشکی بیرجند به تصویب رسیده است.

References

1. Ahada CP, Suthar S. Groundwater nitrate contamination and associated human health risk assessment in southern districts of Punjab, India. Environmental science and pollution research. 2018;25:25336-47.
2. Chen J, Wu H, Qian H, Gao Y. Assessing nitrate and fluoride contaminants in drinking water and their health risk of rural residents living in a semiarid region of Northwest China. Exposure and Health. 2017;9:183-95.
3. Nakagawa K, Amano H, Takao Y, Hosono T, Berndtsson R. On the use of coprostanol to identify source of nitrate pollution in groundwater. Journal of Hydrology. 2017;550:663-8.
4. Li Z, Yang Q, Xie C, Lu X. Source identification and health risks of nitrate contamination in shallow groundwater: a case study in Subei Lake basin. Environmental Science and Pollution Research. 2023;30(5):13660-70.
5. Pouye A, Cissé Faye S, Diédhiou M, Gaye CB, Taylor RG. Nitrate contamination of urban groundwater and heavy rainfall: Observations from Dakar, Senegal. Vadose Zone Journal. 2023:e20239.
6. Iqbal J, Su C, Wang M, Abbas H, Baloch MYJ, Ghani J, et al. Groundwater fluoride and nitrate contamination and associated human health risk assessment in South Punjab, Pakistan. Environmental Science and Pollution Research. 2023;30(22):61606-25.
7. El Bilali A, Taleb A, Brouziyne Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. Agricultural Water Management. 2021;245:106625.
8. Tung TM, Yaseen ZM. A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology. 2020;585:124670.
9. Lu H, Ma X. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere. 2020;249:126169.
10. Castrillo M, García ÁL. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. Water Research. 2020;172:115490.
11. Meyers G, Kapelan Z, Keedwell E. Short-term forecasting of turbidity in trunk main networks. Water research. 2017;124:67-76.
12. Bilali AE, Taleb A, Mazigh N, Mokhliss M. Prediction of chemical water quality used for drinking purposes based on artificial neural networks. Moroccan Journal of Chemistry. 2020;8(3):8-3 (2020) 665-672.
13. Fijani E, Barzegar R, Deo R, Tziritis E, Skordas K. Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters. Science of the total environment. 2019;648:839-53.
14. Wagh VM, Panaskar DB, Muley AA, Mukate SV, Lolage YP, Aamalawar ML. Prediction of groundwater suitability for irrigation using artificial neural network model: a case study of Nanded tehsil, Maharashtra, India. Modeling Earth Systems and Environment. 2016;2:1-10.
15. Ahmed SH, Abed MF. Evaluating Groundwater Quality for Sustainable Drinking and Irrigation Purposes and Assessing Nitrate Risks on Human Health in Rural Areas. 2020.
16. Hosseinzadeh A, Zhou JL, Altaee A, Baziar M, Li X. Modeling water flux in osmotic membrane bioreactor by adaptive network-based fuzzy inference system and artificial neural network. Bioresource technology. 2020;310:123391.
17. Zhu X, Wang X, Ok YS. The application of machine learning methods for prediction of metal sorption onto biochars. Journal of hazardous materials. 2019;378:120727.
18. Sajedi-Hosseini F, Malekian A, Choubin B, Rahmati O, Cipullo S, Coulon F, et al. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Science of the total environment. 2018;644:954-62.
19. Knoll L, Breuer L, Bach M. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Science of the total environment. 2019;668:1317-27.
20. Benaroussi O, Djellal M. Mapping groundwater vulnerability to nitrate contamination using machine learning techniques 2022.
21. Latif SD, Azmi M, Ahmed AN, Fai CM, El-Shafie A. Application of artificial neural network for forecasting nitrate concentration as a water quality parameter: a case study of Feitsui Reservoir, Taiwan. Int J Des Nat Ecodynamics. 2020;15:647-52.
22. Ubah J, Orakwe L, Ogbu K, Awu J, Ahaneku I, Chukwuma E. Forecasting water quality parameters using artificial neural network for irrigation purposes. Scientific Reports. 2021;11(1):24438.