Journal of Water and Wastewater Science and Engineering (JWWSE)
Vol. 8, No. 3, PP. 25-34, Autumn 2023
DOI: 10.22112/JWWSE.2023.360223.1326

نشریه علمی
علوم و مهندسی آب و فاضلاب
سال هشتم، شماره ۳، صفحات ۲۵-۳۴، پاییز ۱۴۰۲

Research Paper

Investigating the Performance of the Combined Dagging Method with the Hoeffding Tree Base Algorithm in the Qualitative Classification of Drinking Water

Mohammad Taghi Sattari[1*] and Sahar Javidan[2]
1- Associate Professor, Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran.
2- M.Sc. Student, Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran.
* Corresponding Author, Email: mtsattar@tabrizu.ac.ir

Received: 11/09/2022
Revised: 11/04/2023
Accepted: 26/04/2023

Abstract

For the effective qualitative management of drinking water, it is necessary to estimate the level of water pollution. In this research, to calculate the quality index of drinking water from the chemical parameters of Total Hardness, Alkalinity, Electrical Conductivity, Total Dissolved Solids, Calcium, Sodium, Magnesium, Potassium, Chlorine, Carbonate, Bicarbonate, and Sulfate in the hydrometric station of Bagh Kelayeh, Qazvin province used in the statistical period of 23 years (1998-2020). According to the calculated numerical values and existing standards, water quality classified into two classes, good and excellent. To predict the quality class of drinking water based on chemical parameters, different combinations of parameters were considered in the form of several scenarios. In this regard, correlation and relief algorithms were used to select different scenarios. Hoeffding tree was used as a basic model for classifying water quality based on different combinations of parameters. Also, the performance of the combined Dagging approach in improving the results was evaluated. The results showed that the combined Dagging improves the water quality classification results. Scenario 6 Dagging with Hoeffding tree base algorithm, including HCO_3, Ca, SO_3, TDS, EC and TH parameters, with Kappa = 1, was introduced as the best method which is able to classify test samples correctly.

Keywords: Drinking Water Quality Index, Hoeffding Tree, Kappa Statistic, Rock Curve.

مقاله پژوهشی

بررسی عملکرد روش ترکیبی دگینگ با الگوریتم پایه درخت هوفدینگ در طبقه‌بندی کیفی آب شرب

محمدتقی ستاری[1*] و سحر جاویدان[2]
۱- دانشیار گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران.
۲- دانشجوی کارشناسی ارشد، گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران.
* نویسنده مسئول، ایمیل: mtsattar@tabrizu.ac.ir

تاریخ دریافت: ۱۴۰۱/۰۶/۲۰
تاریخ اصلاح: ۱۴۰۲/۰۱/۲۲
تاریخ پذیرش: ۱۴۰۲/۰۲/۰۶

چکیده

برای مدیریت مؤثر کیفی آب شرب، برآورد سطح آلودگی آب‌های سطحی ضروری است. در پژوهش حاضر، برای محاسبه شاخص کیفی آب شرب از پارامترهای شیمیایی 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. با توجه به مقادیر عددی محاسبه شده و استانداردهای موجود، کیفیت آب در دو کلاس خوب و عالی طبقه‌بندی شد. برای طبقه‌بندی کلاس کیفی آب شرب براساس پارامترهای شیمیایی، ترکیب‌های مختلفی از پارامترها در قالب چندین سناریو درنظر گرفته شد. در این راستا، برای انتخاب سناریوهای مختلف، از دو روش همبستگی و الگوریتم رلیف استفاده شد. درخت هوفدینگ به‌عنوان مدل پایه برای طبقه‌بندی کلاس کیفی آب براساس ترکیب‌های مختلفی از پارامترهای شیمیایی به‌کار برده شد. هم‌چنین عملکرد روش ترکیبی Dagging در بهبود نتایج، مورد ارزیابی قرارگرفت. نتایج نشان داد که روش ترکیبی Dagging باعث بهبود نتایج طبقه‌بندی کلاس کیفی آب می‌شود. سناریوی ۶ روش Dagging با الگوریتم پایه درخت هوفدینگ، شامل پارامترهای HCO_3، Ca، SO3، TDS، EC و TH، با Kappa = 1، به‌عنوان بهترین روش معرفی شد. این روش توانست تمام نمونه‌های آزمایشی را به‌صورت صحیح، طبقه‌بندی کند.

کلمات کلیدی: آماره کاپا، روش ترکیبی Dagging، درخت هوفدینگ، شاخص کیفی آب شرب، منحنی راک.

۱- مقدمه

با توجه به این‌که نیازهای کشاورزی، صنعتی و خانگی افزایش یافته، در نتیجه رقابت فزاینده‌ای برای تأمین آب پاک نیز به‌وجود آمده است. سیستم‌های آب سطحی به‌دلیل افزایش فشارهای ناشی از گسترش شهری و تغییر کاربری اراضی شهری با اختلالات قابل‌توجهی از طریق احیا، تغییر و آلودگی مواجه هستند. به‌هر طریقی، بسیاری از مشکلات مربوط به آب (مانند سیل یا خشکسالی و آلودگی جدی آب) نتیجه توسعه نامنظم یا نادرست کاربری زمین است (Kavita and Jagdish, 2012). در دو دهه اخیر، آلودگی رودخانه‌های آب به‌عنوان یکی از دغدغه‌های جهانی جوامع، مطرح شده است که توجه کامل محققان محیط‌زیست را می‌طلبد (Kotti et al., 2005). مطالعات بسیاری در زمینه برآورد کیفیت آب سطحی انجام شده است. بگینگ، تقویت و دگینگ، روش‌های نمونه‌گیری مجدد شناخته شده‌ای هستند که با استفاده از الگوریتم یادگیری یکسان برای طبقه‌بندی‌کننده‌های پایه، تنوعی از طبقه‌بندی‌کننده‌ها را مخصوصاً در شرایطی که داده‌ها دارای نویز هستند، تولید و ترکیب می‌کنند (Kotsianti and Kanellopoulos, 2007).

Kotsianti and Kanellopoulos (2007) بر روی یک سری از داده‌های معیار استاندار با استفاده از روش رأی‌گیری از گروه‌های بگینگ، تقویت و دگینگ با ۸ طبقه‌بندی فرعی مدل را پیاده نمودند. در مقایسه‌ای که با گروه‌های بسته‌بندی، تقویت‌کننده و دگینگ ساده با ۲۵ طبقه‌بندی فرعی و همچنین سایر روش‌های ترکیبی معروف، بر روی مجموعه داده‌های معیار استاندارد انجام دادند، به این نتیجه رسیدند که الگوریتم پیشنهادی در اکثر موارد از دقت بیشتری برخوردار بوده است. Babar and Babar (2017) شاخص کیفیت آب رودخانه یامونا را با استفاده از تکنیک‌های داده‌کاوی[۱]، پیش‌بینی کردند. نتایج مطالعه آن‌ها نشان داد که طبقه‌بندی‌کننده‌های درخت تصمیم[۲] و ماشین‌های بردار پشتیبان[۳]، بهترین مدل‌های پیش‌گوینده با نرخ خطای ۰٪ در طول استخراج خودکار کلاس کیفیت آب هستند. در صورتی‌که داده‌های ارائه‌شده نمایشی واقعی از دانش حوزه باشند، تکنیک‌های داده‌کاوی پتانسیل پیش‌بینی سریع کلاس کیفیت آب را خواهند داشت.

ستاری و همکاران (۱۳۹۶) از روش‌های داده‌کاوی برای پیش‌بینی کیفیت آب‌های سطحی رودخانه‌های دامنه شمالی سهند استفاده کردند. آن‌ها نتیجه گرفتند که مدل تصمیم‌گیری درختی با استفاده از چهار پارامتر هدایت الکتریکی[۴]، PH، نسبت جذب سدیم[۵] و سدیم قادر است با دقت زیادی کیفیت آب را طبقه‌بندی کند. دزفولی و همکاران (۱۳۹۶) طبقه‌بندی کیفی آب رودخانه کارون را براساس حداقل پارامترهای کیفی انجام دادند. نتایج مطالعه آن‌ها نشان داد که روش شبکه عصبی احتمالی[۶] با استفاده از پارامترهای کیفی کدورت، کلی‌فرم مدفوعی و کل مواد جامد با دقت ۹۰/۷۸ می‌تواند به طبقه‌بندی کیفی آب بپردازد. Gakii and Jepkoech (2019) با استفاده از مدل درخت تصمیم کیفیت آب در کنیا را طبقه‌بندی و تجزیه و تحلیل کردند. آن‌ها درخت تصمیم J48 و Decision Stump را به‌ترتیب به‌عنوان مدلی با بیشترین و کمترین دقت معرفی کردند. آن‌ها دریافتند که تجزیه و تحلیل قلیائیت آب، سطح pH و هدایت الکتریکی می‌تواند نقش مهمی در ارزیابی کیفیت آب داشته باشد.

Meddouri et al. (2021) با استفاده از مجموعه داده‌های شناخته شده از مخزن یادگیری ماشین UCI عملکرد الگوریتم دگینگ را ارزیابی کردند. برای هر مجموعه داده، به‌ترتیب تعداد نمونه‌ها، تعداد ویژگی‌های عددی (قبل از گسسته‌سازی)، تعداد ویژگی‌های اسمی (پس از گسسته‌سازی)، تعداد کلاس‌ها و تنوع داده‌ها بررسی شد. نتایج نشان داد به‌علت این‌که روش‌های دگینگ از اکثریت رأی برای ترکیب طبقه‌بندی‌کننده استفاده می‌کنند، بنابراین از دقت بیشتری برخوردار هستند. ایشان پیشنهاد نمودند که استفاده از الگوریتم دگینگ می‌تواند برای بهبود عملکرد طبقه‌بندی مجموعه داده مفید باشد. Islam Khan et al. (2021) براساس رگرسیون مؤلفه اصلی[۷] و رویکرد طبقه‌بندی‌کننده تقویت گرادیان[۸]، کیفیت آب را طبقه‌بندی و پیش‌بینی نمودند. آن‌ها رویکرد طبقه‌بندی‌کننده تقویت گرادیان را با دقت طبقه‌بندی ۱۰۰ درصد، به‌عنوان روش برتر معرفی کردند. همچنین دریافتند که این روش، در مقایسه با مدل‌های پیشرفته، عملکرد مطلوبی داشته است.

Sattari et al. (2021) کیفیت آب رودخانه آلاداغ در ترکیه را به‌منظور مصارف آبیاری و کشاورزی طبقه‌بندی کردند. پس از ارزیابی روش‌های داده‌کاوی، نتیجه گرفتند که این روش‌ها، دقت و عملکرد خوبی در طبقه‌بندی کیفیت آب داشتند. هم‌چنین، به‌طور کلی دریافتند که از میان هسته‌های مورد استفاده برای طبقه‌بندی در روش بردار پشتیبان، هسته جهانی پیرسون و از درخت‌های تصمیم، درخت REP[۹] بهترین نتایج را برای طبقه‌بندی کیفیت آب رودخانه آلاداغ داشت. آن‌ها روش درخت هوفدینگ را هم به‌کار بردند که نتایج مطلوبی را به‌دنبال نداشت. Yusri et al. (2022) طبقه‌بندی کیفیت آب را با استفاده از روش ماشین بردار پشتیبان و افزایش گرادیان شدید[۱۰] (XGBoost) انجام دادند.

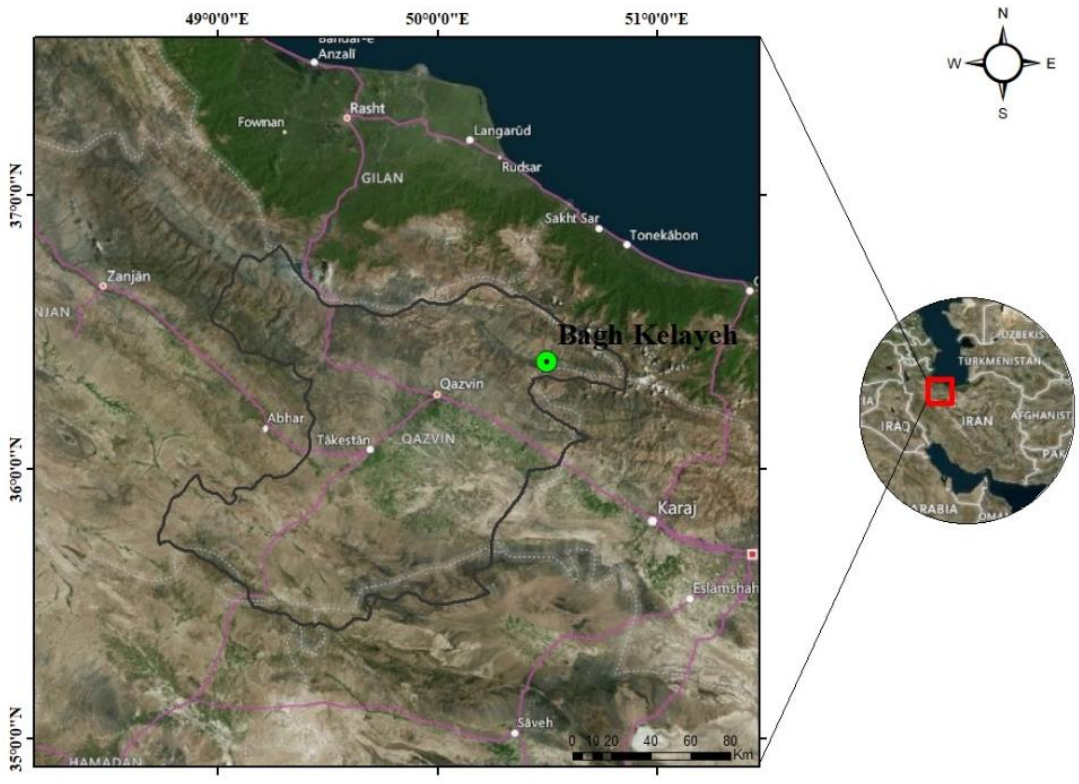
نتایج تحقیق آن‌ها نشان داد که مدل XGBoost با دقت ۹۴ درصد، بهتر از مدل SVM عمل کرده و تنها ۶ درصد خطا در طبقه‌بندی کیفیت آب داشته است.

با توجه به هزینه‌های سرسام‌آور منیتورینگ و نمونه‌برداری از رودخانه‌ها برای تعیین کیفیت آب شرب، ارائه مدل‌هایی که بتواند با حداقل تعداد پارامتر هیدروشیمیایی با دقت نسبتاً قابل‌قبولی کلاس کیفی آب را مشخص کند، یکی از اولویت‌های مدیران کیفی آب است. در این راستا هدف پژوهش حاضر، طبقه‌بندی کلاس کیفی آب شرب، با استفاده از مدل‌های داده‌کاوی براساس داده‌های مشاهداتی مربوط به پارامترهای کیفی آب ایستگاه هیدرومتری باغ کلایه در استان قزوین است. با توجه به این‌که تاکنون از روش دگینگ در زمینه مدل سازی کیفی آب استفاده نشده است، بررسی تأثیر استفاده از این روش در بهبود نتایج، یکی از نوآوری‌های پژوهش حاضر است.

۲- مواد و روش‌ها

۲-۱- منطقه مورد مطالعه و داده‌های مورد استفاده

باغ کلایه، روستایی از توابع بخش رودبار الموت شهرستان قزوین در استان قزوین است. ایستگاه هیدرومتری باغ کلایه دارای عرض جغرافیایی ″۳۸ ′۲۳ °۳۶، طول جغرافیایی ″۵۱ ′۲۹ °۵۰ بوده و در ارتفاع ۱۲۸۷ متر از سطح دریا واقع شده است. استان قزوین در بخش شمال غربی کشور ایران واقع شده و مساحت آن حدود km^2 ۱۵۸۲۰۰ است. میانگین بارش سالانه در ۲۰ سال اخیر mm ۳۰۲/۵۱ و متوسط دمای سالانه °C ۱۵/۲۱ بوده که براساس طبقه‌بندی دومارتن دارای اقلیم نیمه‌خشک و براساس اقلیم‌نمای آمبروژه دارای اقلیم خشک سرد است (جاویدان و همکاران، ۱۴۰۱). شکل ۱ موقعیت مکانی ایستگاه موردمطالعه را نشان می‌دهد.



شکل ۱- موقعیت مکانی ایستگاه هیدرومتری باغ کلایه

در پژوهش حاضر برای محاسبه شاخص کیفی آب[11] (WQI) و طبقه‌بندی کلاس کیفی آب شرب ایستگاه هیدرومتری باغ کلایه، از پارامترهای کیفی شامل سختی کل (TH)، قلیائیت (pH)، هدایت الکتریکی (EC)، کل مواد جامد محلول[12] (TDS)، کلسیم (Ca)، سدیم (Na)، منیزیم (Mg)، پتاسیم (K)، کلر (Cl)، کربنات (CO_3)، بی‌کربنات (HCO_3) و سولفات (SO_4) استفاده شد. کلاس کیفی به‌دست آمده از این‌روش، به‌عنوان خروجی‌های هدف در فرآیند مدل‌سازی مورد استفاده قرارگرفتند. دوره آماری در مطالعه حاضر، ۲۳ ساله (۱۹۹۸-۲۰۲۰) درنظر گرفته شد. مشخصات آماری متغیرهای مورد استفاده در جدول ۱ ارائه شده است.

جدول ۱- مشخصات آماری پارامترهای مورد استفاده

Statistic	حداقل	حداکثر	میانگین	واریانس
سختی کل (میلی‌گرم بر لیتر)	۹۵/۰۰	۴۸۱/۵۰	۲۷۸/۱۵	۳۴۰۷/۳۸
قلیاییت	۴/۵۰	۸/۴۰	۷/۸۳	۰/۱۱
هدایت الکتریکی (میکروموس بر سانتی‌متر)	۲۷۹/۰۰	۱۰۴۸/۰۰	۶۲۷/۹۴	۱۷۴۸۰/۸۷
کل مواد جامد محلول (میلی‌گرم بر لیتر)	۱۸۶/۰۰	۶۶۳/۰۰	۳۸۸/۰۵	۵۹۸۳/۸۱
کلسیم (میلی اکی‌والان بر لیتر)	۰/۰۰	۱۵۹/۸۰	۷۳/۹۴	۳۸۰/۲۴
سدیم (میلی اکی‌والان بر لیتر)	۰/۴۶	۶۰/۴۹	۱۷/۲۶	۸۲/۹۸
منیزیم (میلی اکی‌والان بر لیتر)	۲/۷۶	۵۸/۲۰	۲۲/۱۸	۰/۹۳۷۰
پتاسیم (میلی اکی‌والان بر لیتر)	۰/۳۹	۱۹/۵۰	۱/۹۹	۲/۷۹
کلر (میلی اکی‌والان بر لیتر)	۰/۰۰	۸۹/۶۰	۲۷/۳۶	۱۹۸/۶۱
کربنات (میلی اکی‌والان بر لیتر)	۰/۰۰	۳۳/۰۰	۰/۲۱	۴/۱۷
سولفات (میلی اکی‌والان بر لیتر)	۲۲/۰۸	۳۷۴/۸۸	۱۴۰/۲۱	۲۱۲۹/۱۹
بی کربنات (میلی اکی‌والان بر لیتر)	۵۰/۰۲	۳۹۱/۶۲	۱۶۳/۶۰	۱۵۶۵/۲۱

۲-۲- سناریوهای مورد استفاده

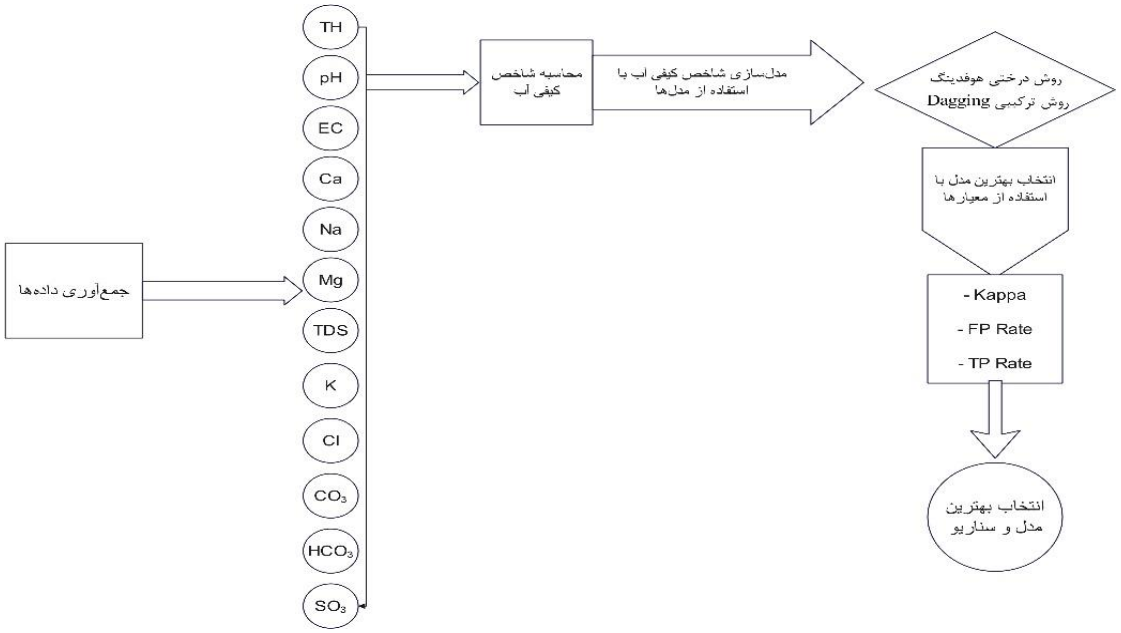
در این پژوهش برای انجام مدل‌سازی با روش‌های داده‌کاوی از سناریوهای مختلفی استفاده شد تا بهترین ترکیب ورودی مشخص شود و تأثیر انتخاب پارامترهای شیمیایی مختلف مورد بررسی قرارگیرد. براین اساس سناریوهای استفاده شده در پژوهش حاضر در جدول ۲ ارائه شده است. انتخاب ترکیب‌های ورودی مختلف، با استفاده از روش همبستگی و الگوریتم رلیف[۱۳] انجام شد. استفاده از الگوریتم رلیف که برای کاهش ابعاد مسئله مورد استفاده قرار می‌گیرد؛ توسط Kira and Rendell (1992) پیشنهاد شده است. مزیت‌های استفاده از این روش، ساده بودن اصول و عدم پیچیدگی آن، قابل‌حل بودن با توابع چندجمله‌ای مرتبه پایین، قابل استفاده بودن برای داده‌های پیوسته و نیاز به تعداد کم داده‌های آموزشی درنظر گرفته می‌شود. 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 (۱،۰) قرارگیرد. الگوریتم مذکور، m بار تکرار شده و در هر مرتبه از یک بردار وزنی متفاوت که از صفر شروع می‌شود، 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 موردنظر را توسط تابع فاصله اقلیدسی انتخاب می‌کند. پس از m تکرار، هر یک از عناصر بردار وزن توسط m تقسیم‌بندی می‌شوند. نتیجه این عمل این است که یک بردار مرتبط به‌دست می‌آید. چنان‌چه مقدار بردار مرتبط یک ویژگی، از آستانه تعریف شده بیشتر شود، آن ویژگی انتخاب می‌شود. این الگوریتم نه به‌عنوان یک روش پیش‌پردازش است نه روش طبقه‌بندی.

الگوریتم رلیف به‌منظور مشخص نمودن تأثیر متغیرهای مستقل بر روی متغیر پاسخ و یا هدف به‌کار می‌رود. الگوریتم رلیف ترکیب‌های متفاوتی از متغیرهای ورودی در راستای ایجاد سناریوهای مختلف مدل‌سازی را ارائه می‌کند (Kira and Rendell, 1992). انتخاب ویژگی بر اساس همبستگی، یک روش معمول و پرکاربرد برای انتخاب متغیرهای ورودی و کاهش ابعاد مسئله است. روش همبستگی به زیرمجموعه‌هایی که دارای ویژگی‌هایی با بیشترین ضریب همبستگی با کلاس نمونه مورد نظر هستند، امتیاز می‌دهد و متغیرهایی که بیشترین امتیاز را دارا باشند، به‌عنوان متغیر اصلی درنظر می‌گیرد. این الگوریتم توانایی بالایی در تشخیص سریع داده‌های نامربوط، اضافی و دارای خطا دارد که عموماً منجر به حذف نیمی از داده‌ها می‌شود. این ویژگی با کاهش ابعاد مسئله سبب افزایش بهره‌وری مدل‌ها می‌شود (Hall, 1999). از بین داده‌های مورد مطالعه، ۷۰٪ برای واسنجی و ۳۰٪ برای صحت‌سنجی انتخاب شدند. برای طبقه‌بندی کلاس کیفی آب شرب، از روش درخت هوفدینگ[۹] استفاده شد و عملکرد روش Dagging با الگوریتم پایه درخت هوفدینگ مورد بررسی قرار گرفت.

۲-۳- روش‌های مورد مطالعه

در این پژوهش برای تعیین کیفیت آب از دیدگاه شرب، ابتدا شاخص کیفی آب به‌عنوان یک شاخص عددی با استفاده از پارامترهای شیمیایی مؤثر در کیفیت آب محاسبه شد. سپس با استفاده از استانداردهای موجود کلاس و طبقه کیفی آب از نقطه‌نظر مساعد بودن و یا نبودن برای شرب تعیین شد. از آن‌جایی که نمونه‌برداری از منبع آب برای تعیین کیفیت آب زمان‌بر و هزینه‌بر است، لذا به‌کمک روش‌های داده‌کاوی سعی شد با استفاده از حداقل تعداد پارامتر شیمیایی شاخص کیفی آب و به‌تبع آن کلاس کیفی آب مدل‌سازی شود. در همین راستا

ترکیب‌های مختلفی از پارامترهای شیمیایی در قالب چندین سناریو به‌عنوان ورودی مدل‌های داده‌کاوی براساس نمونه‌های مشاهداتی درنظر گرفته شد. در همین سناریوها کلاس و طبقه کیفی آب که براساس شاخص کیفی آب برای نمونه‌های مشاهداتی قبلاً تعیین شده بود به‌عنوان خروجی مدل‌ها برای آموزش و انجام فرآیند مدل‌سازی درنظر گرفته شد. کارکرد کلی هریک از مدل‌ها و الگوریتم‌های داده‌کاوی مورد استفاده در این پژوهش معطوف به افزایش دقت مدل‌ها در پیش‌بینی شاخص کیفی آب بود. با استفاده از معیارهای ارزیابی نتایج هریک از مدل‌ها و الگوریتم‌ها برای سناریوهای معرفی شده مورد مقایسه قرارگرفت و نهایتاً بهترین سناریو (ترکیب پارامترهای ورودی) و دقیق‌ترین مدل و الگوریتم داده‌کاوی انتخاب شد. در شکل ۲ روند انجام تحقیق ارائه شده است.



شکل ۲- روندنمای مراحل انجام تحقیق

۲-۳-۱- شاخص کیفیت آب شرب

از فرمول‌های (۱) تا (۳) برای محاسبه شاخص کیفی آب و طبقه‌بندی کلاس کیفی، استفاده شد (Singh, 1992).

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

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

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

که w: وزن مربوط به هر پارامتر با توجه به اهمیت آن در شرب، 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 است.

پس از محاسبه مقادیر WQI، کلاس کیفی آب شرب به پنج دسته آب عالی، خوب، بد، بسیار ضعیف و نامناسب برای آشامیدن تقسیم می‌شود (جدول ۲).

جدول ۲- طبقه بندی کیفیت آب بر اساس ارزش WQI

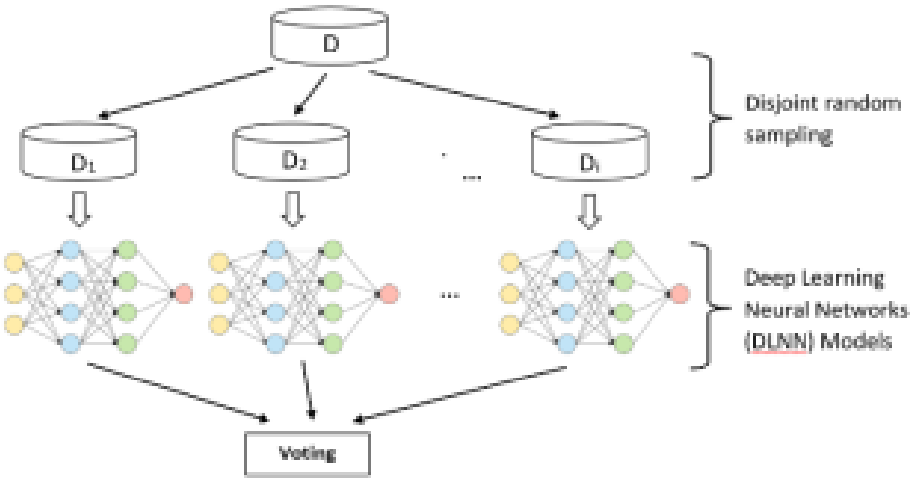
طبقه‌بندی کیفیت آب آشامیدنی		
نوع آب	کلاس	محدوده شاخص WQI
آب عالی	I	کمتر از ۵۰
آب خوب	II	۵۰-۱۰۰
آب بد	III	۱۰۰-۲۰۰
آب بسیار بد	IV	۲۰۰-۳۰۰
آب نامناسب برای آشامیدن	V	بیشتر از ۳۰۰

۲-۳-۲- روش درختی هوفدینگ

مدل درختی هوفدینگ[۱۴] که به‌عنوان یک الگوریتم افزایشی معروف درخت تصمیم شناخته می‌شود، در ابتدا برای رفع مشکلات طبقه‌بندی در استخراج جریان داده در مقیاس بزرگ پیشنهاد شد (Domingos and Hulten, 2003). در روش مدل HT، هر نمونه در پایگاه داده آموزشی تنها یک‌بار اسکن می‌شود. از این‌رو، این مدل دارای راندمان محاسباتی برجسته با رم نسبتاً پایین است. علاوه‌بر این، عملیات طبقه‌بندی را می‌توان در زمانی که مدل HT در حال رشد است، اجرا کرد که به‌وضوح با مدل‌های DT معمولی متفاوت است (Mehta and Sanghavi, 2019).

۲-۳-۳- روش Dagging

این طبقه‌بندی‌کننده، تعدادی چین‌های منفصل و طبقه‌بندی‌شده از داده‌ها ایجاد می‌کند و هر تکه از داده‌ها را به‌یک کپی از طبقه‌بندی‌کننده پایه ارائه‌شده، تغذیه می‌کند. پیش‌بینی‌ها از طریق اکثریت آرا انجام می‌شود. نمونه‌گیری غیرمجاز تصادفی و ناهمگن در مجموعه داده آموزشی اعمال می‌شود. پس از نمونه‌برداری مجدد از مجموعه داده آموزشی، یک طبقه‌بندی پایه برای هر نمونه ایجاد می‌شود. سپس نتایج این طبقه‌بندی‌کننده‌های چندگانه با استفاده از رأی اکثریت ترکیب می‌شوند. بنابراین روش ترکیبی Dagging یک مجموعه موازی درنظر گرفته می‌شود (Ting and Witten, 1997). شکل ۳ نمای کلی از این روش را ارائه می‌دهد.



شکل ۳- نمای کلی روش ترکیبی Dagging (Elish and Elish, 2021)

۲-۴- معیارهای ارزیابی

در هر پژوهش مبتنی بر روش‌های داده‌کاوی، از آن‌جایی‌که روش‌ها در رقابت بایکدیگر عمل می‌کنند، ضروری است تا با به‌کار بردن معیارهایی دقت مدل‌های مورد استفاده را تعیین نموده و بر این اساس نسبت به اولویت‌بندی مدل‌ها و انتخاب مدل برتر اقدام نمود. اساساً در شرایطی که خروجی مدل‌ها کمی (عددی) و یا کیفی (غیرعددی) باشند، معیارهای ارزیابی متفاوتی برای سنجش دقت مدل‌ها به‌کار می‌رود. در این پژوهش خروجی مدل‌های داده‌کاوی کیفی (غیرعددی) هستند و بر این اساس معیارهای ارزیابی، معطوف به مقایسه تعداد صحیح موارد پیش‌بینی شده در برابر تعداد غیرصحیح موارد پیش‌بینی شده در فرآیند مدل‌سازی و طبقه‌بندی است.

برای مقایسه کلاس کیفی به‌دست آمده از روش‌های داده‌کاوی با کلاس کیفی محاسبه شده از WQI از آماره کاپا، نرخ مثبت صحیح[۱۵] (TPR) و نرخ مثبت غلط[۱۶] (FPR) استفاده شد. فرمول‌های آماره‌های فوق به ترتیب در روابط (۴) تا (۶) ارائه شده است.

$$kappa = p_i = (PA_0 - PA_E)/(1 - PA_E) \quad (۴)$$

$$Sensitivity = TPR = \frac{TP}{TP + FN} \quad (۵)$$

$$1 - Specificity = FPR = \frac{TN}{TN + FP} \quad (۶)$$

که PA_0: میزان توافق دو ارزیاب، PA_E: میزان توافق مورد انتظار، TP: مثبت صحیح، FN: منفی غلط، FP: مثبت غلط و TN: منفی صحیح هستند.

منحنی راک برای مقایسه دو ویژگی عملیاتی نرخ مثبت واقعی (TPR) یا حساسیت و نرخ مثبت کاذب (FPR) یا ویژگی به‌کار می‌رود. هم‌چنین به‌عنوان منحنی مشخصه عملکرد گیرنده شناخته می‌شود. منحنی مشخصه عملکرد گیرنده، یک نمایش گرافیکی است که عملکرد الگوریتم‌های طبقه‌بندی‌کننده داده‌کاوی را تفسیر می‌کند. این نتیجه ترسیم نرخ مثبت واقعی در مقابل نرخ مثبت کاذب در تنظیمات آستانه‌های مختلف است. TPR در امتداد محور y و FPR در امتداد محور x رسم می‌شود. عملکرد هر الگوریتم به‌عنوان یک نقطه در منحنی ROC نشان داده می‌شود.

۳- نتایج و بحث

در پژوهش حاضر برای برآورد کلاس کیفی آب با استفاده از شاخص 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 به‌کار برده شد. پارامترهای ورودی سناریوهای مورد مطالعه در جدول ۳ ارائه شده است.

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

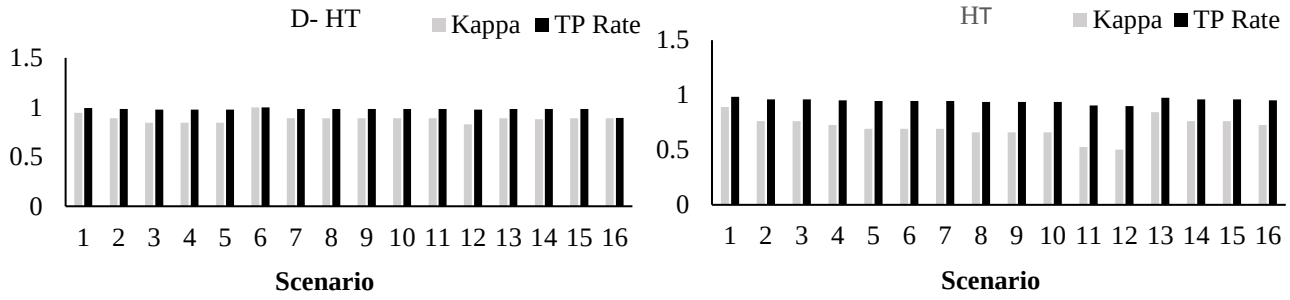
شماره سناریو	ورودی‌ها	روش انتخاب سناریو
۱	TH	ماتریس همبستگی
۲	TH, EC	
۳	TH, EC, TDS	
۴	TH, EC, TDS, SO_4	
۵	TH, EC, TDS, SO_4, Ca	
۶	TH, EC, TDS, SO_4, Ca, HCO_3	
۷	TH, EC, TDS, SO_4, Ca, HCO_3, Mg	
۸	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl	
۹	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na	
۱۰	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K	
۱۱	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K, CO_3	
۱۲	TH, EC, TDS, SO_4, Ca, HCO_3, Mg, Cl, Na, K, CO_3, PH	
۱۳	TH, K	الگوریتم رلیف
۱۴	TH, K, SO_4	
۱۵	TH, K, SO_4, TDS	
۱۶	TH, K, SO_4, TDS, EC	

برای برآورد مقادیر کیفی WQI از روش درختی هوفدینگ و روش ترکیبی Dagging استفاده شد. پس از طبقه‌بندی کلاس کیفی آب در ایستگاه مورد مطالعه، مشخص شد که تمام نمونه‌های برداشت شده در کلاس کیفی یک و دو، یعنی کیفیت عالی (۱۳/۲۷ درصد) و خوب (۸۶/۷۳ درصد) قرارگرفتند. نتایج مدل‌سازی کلاس کیفی آب در جدول ۴ ارائه شده است.

جدول ۴- معیارهای ارزیابی برای برآورد مقادیر کیفی WQI

سناریو	روش									
	روش درختی هوفدینگ					روش ترکیبی Dagging				
	Kappa	RMSE	TP Rate	FP Rate	ROC Area	Kappa	RMSE	TP Rate	FP Rate	ROC Area
۱	۰/۸۹۱	۰/۱۱۷	۰/۹۸۴	۰/۰۰۱	۱/۰۰۰	۰/۹۴۳	۰/۱۰۰	۰/۹۹۲	۰/۰۰۱	۱/۰۰۰
۲	۰/۷۶۲	۰/۱۶۷	۰/۹۶۱	۰/۰۰۳	۱/۰۰۰	۰/۸۹۱	۰/۱۰۴	۰/۹۸۴	۰/۰۰۱	۱/۰۰۰
۳	۰/۷۶۲	۰/۱۸۶	۰/۹۶۱	۰/۰۰۳	۱/۰۰۰	۰/۸۴۵	۰/۱۰۹	۰/۹۷۶	۰/۰۰۲	۱/۰۰۰
۴	۰/۷۲۶	۰/۱۹۸	۰/۹۵۳	۰/۰۰۴	۱/۰۰۰	۰/۸۴۵	۰/۱۱۵	۰/۹۷۶	۰/۰۰۲	۱/۰۰۰
۵	۰/۶۹۲	۰/۲۱۰	۰/۹۴۵	۰/۰۰۴	۱/۰۰۰	۰/۸۴۵	۰/۱۲۲	۰/۹۷۶	۰/۰۰۲	۱/۰۰۰
۶	۰/۶۹۲	۰/۲۱۱	۰/۹۴۵	۰/۰۰۴	۱/۰۰۰	۱/۰۰۰	۰/۰۹۳	۱/۰۰۰	۰/۰۰۰	۱/۰۰۰
۷	۰/۶۹۲	۰/۲۰۱	۰/۹۴۵	۰/۰۰۴	۱/۰۰۰	۰/۸۹۱	۰/۱۰۱	۰/۹۸۴	۰/۰۰۱	۱/۰۰۰
۸	۰/۶۶۱	۰/۲۲۷	۰/۹۳۷	۰/۰۰۵	۰/۹۹۸	۰/۸۹۱	۰/۱۳۲	۰/۹۸۴	۰/۰۰۱	۰/۹۹۷
۹	۰/۶۶۱	۰/۲۳۸	۰/۹۳۷	۰/۰۰۵	۰/۹۹۸	۰/۸۹۱	۰/۱۳۰	۰/۹۸۴	۰/۰۰۱	۰/۹۹۵
۱۰	۰/۶۶۱	۰/۲۴۸	۰/۹۳۷	۰/۰۰۵	۰/۹۹۸	۰/۸۹۱	۰/۱۳۶	۰/۹۸۴	۰/۰۰۱	۰/۹۹۶
۱۱	۰/۵۲۶	۰/۲۹۱	۰/۹۰۶	۰/۱۱۰	۰/۹۲۱	۰/۸۹۱	۰/۱۱۳	۰/۹۸۴	۰/۰۰۱	۰/۹۹۸
۱۲	۰/۵۰۳	۰/۲۹۸	۰/۸۹۸	۰/۱۱۰	۰/۹۲۲	۰/۸۲۹	۰/۱۳۰	۰/۹۷۶	۰/۱۰۴	۰/۹۹۵
۱۳	۰/۸۴۵	۰/۱۳۸	۰/۹۷۶	۰/۰۰۲	۰/۹۹۹	۰/۸۹۱	۰/۱۱۳	۰/۹۸۴	۰/۰۰۱	۰/۹۹۷
۱۴	۰/۷۶۲	۰/۱۵۴	۰/۹۶۱	۰/۰۰۳	۰/۹۹۷	۰/۸۸۰	۰/۱۰۵	۰/۹۸۴	۰/۱۰۴	۰/۹۹۶
۱۵	۰/۷۶۲	۰/۱۷۶	۰/۹۶۱	۰/۰۰۳	۱/۰۰۰	۰/۸۹۱	۰/۰۹۱	۰/۹۸۴	۰/۰۰۱	۱/۰۰۰
۱۶	۰/۷۲۶	۰/۱۹۸	۰/۹۵۳	۰/۰۰۴	۱/۰۰۰	۰/۸۹۱	۰/۰۹۰	۰/۸۹۴	۰/۰۰۱	۱/۰۰۰

همان‌گونه که از جدول ۴ مشخص است، استفاده از روش ترکیبی Dagging باعث بهبود نتایج درخت هوفدینگ شده است. در روش HT سناریو اول شامل TH بالاترین دقت را دارد. با بهره‌گیری از روش Dagging مقدار کاپا در سناریوی اول از ۰/۸۹۱ به ۰/۹۴۳ افزایش یافته است. در حالت کلی از بین دو روش و ۱۶ سناریوی مورد مطالعه، سناریوی ۶ روش Dagging با الگوریتم پایه HT شامل $TH, EC, TDS, SO_4, Ca, HCO_3$ و Kappa = 1، RMSE = 0.093، TPR = 1، FPR = 0 و ROC Area = 1 بهترین عملکرد را داشته است. شکل ۴ نمودار ستونی مقادیر Kappa و TPR روش‌های HT و D- HT را نشان می‌دهد.



شکل ۴- نمودار ستونی مقادیر کاپا و نرخ مثبت صحیح روش‌های مورد مطالعه

برتر بودن سناریوی ۱ در روش HT و سناریوی ۶ در روش D- HT از نمودارهای شکل ۴ نیز مشهود است. جدول‌های ۵ و ۶ ماتریس اغتشاش به‌دست آمده از نتایج سناریوهای برتر هر دو روش مورد مطالعه را نشان می‌دهد. این جدول‌ها نحوه توزیع نمونه‌ها در بین کلاس‌های مختلف را نشان می‌دهد. چنان‌چه نمونه‌ای در این جدول‌ها خارج از قطر اصلی قرارگیرد، به این معنی است که در کلاس اشتباهی طبقه‌بندی شده است. به‌عبارت دیگر، هر چه تعداد نمونه‌ها در قطر اصلی بیشتر باشد، دقت و کارایی مدل در تخمین کلاس موردنظر، بیشتر می‌شود.

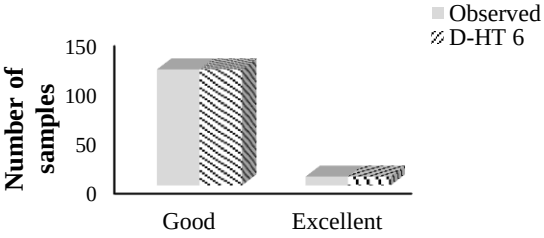
جدول ۵- ماتریس اغتشاش سناریو برتر روش HT

		طبقه‌بندی شده با مدل	
		عالی	خوب
واقعی	خوب	۲	۱۱۶
	عالی	۹	۰

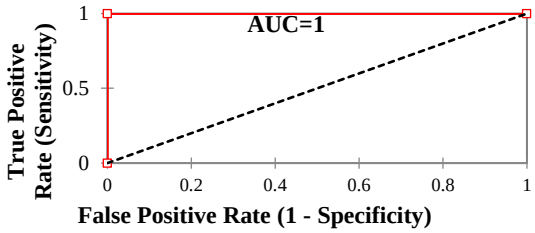
جدول ۶- ماتریس اغتشاش سناریو برتر روش D- HT

		طبقه‌بندی شده با مدل	
		عالی	خوب
واقعی	خوب	۰	۱۱۸
	عالی	۹	۰

با توجه به جدول ۵، از بین ۱۲۷ نمونه آزمایشی، تنها ۲ نمونه به اشتباه طبقه‌بندی شده است. یعنی به‌جای کلاس کیفی خوب (II)، در کلاس عالی (I) قرارگرفته است. اما در روش D-HT6 که ماتریس اغتشاش آن در جدول ۶ مشخص شده است، تمام نمونه‌ها در کلاس صحیح، طبقه‌بندی شده‌اند. شکل ۵ نیز توزیع سه‌بعدی کلاس کیفی آب برای داده‌های مشاهداتی و مدل D- HT6 برای دوره آزمون را نشان می‌دهد. بنابراین می‌توان نتیجه گرفت که روش Dagging عملکرد خوبی در برآورد مقادیر کیفی WQI داشته و باعث بهبود نتایج شده است. منحنی راک سناریو و روش برتر معرفی شده در پژوهش حاضر در شکل ۶ مشخص شد.



شکل ۵- توزیع سه‌بعدی کلاس کیفی آب برای داده‌های مشاهداتی و مدل D- HT6 برای دوره آزمون



شکل ۶- منحنی راک روش D- HT6

هنگامی‌که حساسیت افزایش پیدا کند، نرخ مثبت کاذب نیز افزایش خواهدیافت. بنابراین در منحنی راک امکان مقایسه نرخ مثبت صحیح و نرخ مثبت کاذب، در هر نقطه بر روی منحنی وجود دارد. سطح زیر منحنی نشان‌دهنده کیفیت مدل‌های مورد مطالعه است. هرچه نرخ مثبت صحیح یا حساسیت بیشتر باشد، بدین معنی است که نمونه‌ها به‌درستی طبقه‌بندی شده‌اند. در نتیجه سطح زیر منحنی بیشتر شده و به ۱۰۰ درصد خواهد رسید. با توجه به توضیحات فوق و شکل ۶، چون سطح زیر منحنی برابر با یک شده است، می‌توان نتیجه گرفت که روش مورد استفاده دقت بالایی داشته است.

۴- نتیجه‌گیری

در پژوهش حاضر، کلاس کیفی آب ایستگاه هیدرومتری باغ کلایه مورد بررسی قرارگرفت. ابتدا مقادیر عددی کیفی شاخص WQI آب براساس داده‌های مشاهداتی شده پارامترهای شیمیایی محاسبه و براساس استانداردهای موجود، کلاس کیفیت آب شرب مشخص شد. برای طبقه‌بندی کیفی 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. طبقه‌بندی کلاس کیفی آب با روش درخت هوفدینگ انجام شد. در مدل‌سازی کیفی، تاثیر روش ترکیبی Dagging با الگوریتم پایه درخت هوفدینگ در بهبود نتایج، مورد ارزیابی قرارگرفت. برای انتخاب روش و سناریو برتر از آماره کاپا و منحنی راک استفاده شد. در مدل‌سازی کیفی بهره‌گیری از روش Dagging باعث افزایش دقت و کاهش خطا شد. در طبقه‌بندی کلاس کیفی آب شرب، سناریوی ۶ روش D-HT شامل پارامترهای TH, EC, TDS, SO_4, Ca, HCO_3 از دقت قابل‌توجهی برخوردار بود و توانست تمام نمونه‌ها را در کلاس کیفی صحیح، طبقه‌بندی کند. بنابراین این روش توانست به‌جای استفاده از ۱۲ پارامتر شیمیایی، تنها با استفاده از ۶ پارامتر، با دقت مناسبی کلاس کیفی تمام نمونه‌های ایستگاه مورد مطالعه را به‌درستی، طبقه‌بندی کند. با توجه به این‌که تمام روش‌های مورد استفاده در طبقه‌بندی کلاس کیفی، دقت قابل‌قبولی داشتند، لذا در صورت کمبود داده و عدم دسترسی به تمام پارامترهای شیمیایی، می‌توان با استفاده از تعداد محدودی از پارامترها و با بهره‌گیری از روش‌های داده‌کاوی، نتایج مناسب و قابل‌قبولی را به‌دست آورده و کلاس کیفیت آب شرب را طبقه‌بندی نمود. نتایج به‌دست آمده نشان داد؛ آب منطقه مورد مطالعه از کیفیت قابل‌قبولی برای شرب برخوردار است.

۵- پی‌نوشت‌ها

1- Data Mining
2- Decision Tree
3- Support Vector Machine
4- Electrical Conductivity
5- Sodium Absorption Ratio
6- Probabilistic Neural Network
7- Principal Component Regression
8-Gradient Boosting Classifier Approach
9- Reduced Error Pruning Tree
10- Extreme Gradient Boosting
11- Water Quality Index
12- Total Dissolved Solids
13- Relief
14- Hoeffding Tree
15- True Positive Rate
16- False Positive Rate

۶- مراجع

جاویدان، س.، ستاری، م. ت.، کریم‌زاده، پ.، و مهرابی، ا.، (۱۴۰۱)، "تحلیل عملکرد روش‌های هیدرولوژیکی و داده-مبنا در برآورد میزان رسوب معلق"، *مجله محیط‌زیست و مهندسی آب*، ۸(۲)، ۴۸۰-۴۶۸، https://doi.org/10.22034/jewe.2021.305599.1632.

دزفولی، د.، موغاری، م.ح.، ابراهیمی، ک.، و عراقی‌نژاد، ش.، (۱۳۹۶)، "تعیین طبقه‌بندی کیفی آب بر اساس حداقل پارامترهای کیفی (مطالعه موردی: رودخانه کارون)"، *محیط‌زیست طبیعی، مجله منابع طبیعی ایران*، ۷۰(۳)، ۵۹۵-۵۸۳، https://doi.org/10.22059/jne.2017.213338.1224.

ستاری، م. ت.، میرعباسی، ر.، و عباسقلی نایب‌زاد، م.، (۱۳۹۶)، "استفاده از داده‌کاوی در پیش‌بینی کیفیت آب‌های سطحی (مطالعهٔ موردی: رودخانه‌های دامنهٔ شمالی سهند)"، *اکوهیدرولوژی*، ۴(۲)، ۴۱۹-۴۰۷، https://doi.org/10.22059/ije.2017.61477.

Babar, R., and Babar, S., (2017), "Predicting river water quality index using data mining techniques", *Environmental Earth Sciences*, 76(504), 1-15, https://doi.org/10.1007/s12665-017-6845-9.

Domingos, P., and Hulten, G., (2003), "A general framework for mining massive data streams", *Journal of Computational and Graphical Statistics*, 12(4), 945-949, https://doi.org/10.1198/1061860032544.

Elish, M., and Elish, K., (2021), "An empirical comparison of resampling ensemble methods of deep learning Neural Networks for cross-project software defect prediction", *International Journal of Intelligent Engineering and Systems*, 14(3), 201-209, https://doi.org/10.22266/ijies2021.0630.18.

Gakii, C., and Jepkoech, J., (2019), "A classification model for water quality analysis using desision tree", *European Journal of Computer Science and Information Technology*, 7(3), 1-8.

Hall, M.A., (1999), "Correlation-based feature selection for machine learning", Ph.D. Thesis, University of Waikato.

Khan, M.S.I., Islam, N., Uddin, J., Islam, S., and Nasir, M.K. (2022). 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, https://doi.org/10.1016/j.jksuci.2021.06.003

Kavita, D., and Jagdish, S., (2012), "Water resources management and water quality, case of Bhopal", *International Conference on Chemical, Ecology and Environmental Sciences*, Bangkok.

Khalil, B., Ouarda, T., and 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, https://doi.org/10.1016/j.jhydrol.2011.05.024.

Kira, K., and Rendell, L. A., (1992), "The Feature Selection Problem: Traditional methods and a new algorithm", *Proceedings of the 10th National Conference on Artificial intelligence*, 129-134.

Kotsianti, S.B., and Kanellopoulos, D., (2007), "Combining bagging, boosting and dagging for classification problems", In: Apolloni, B., Howlett, R.J., Jain, L. (eds.), *Knowledge-Based Intelligent Information and Engineering Systems, KES* 2007, Lecture Notes in Computer Science, Vol. 4693, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74827-4_62.

Kotti, M.E., Vlessidis, A.G., Thanasoulias, N.C., and Evmiridis, N.P., (2005), "Assessment of river water quality in Northwestern Greece", *Water Resources Manage*ment, 19, 77-94, https://doi.org/10.1007/s11269-005-0294-z.

Meddouri, N., Khoufi, H., and Maddouri, M., (2021), "A performant dagging approach of classification based on formal concept", *International Journal of Artificial Intelligence and Machine Learning (IJAIML)*, 11(2), 38-62, http://doi.org/10.4018/IJAIML.20210701.oa3.

Mehta, V., and Sanghavi, V., (2019), "Comparative study of various decision tree methods for data stream mining", In: *3rd International Conference on Information and Communication Technology* (ICICT), Springer International Publishing, pp. 371-379, https://doi.org/10.1007/978-981-13-1165-9_34.

Sattari, M.T., Feizi, H., Colak, M., Ozturk, A., Ozturk, F., and Apaydin, H., (2021), "Surface water quality classification using data mining approaches: Irrigation along the Aladag River", *Irrigation and Drainage*, 70(5), 1227-1246, https://doi.org/10.1002/ird.2594.

Singh, D.F, (1992), "Studies on the water quality index of some major rivers of Pune, Maharashtra", *Proceedings of the Academy of Environmental Biology* 1, 61-66.

Ting, K.M., and Witten, I.H., (1997), "Stacking bagged and dagged models", In: *Fourteenth International Conference on Machine Learning*, San Francisco, CA, pp. 367-375.

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