**پیش‌بینی کیفیت آب با استفاده از مدل‌های یادگیری ماشین مبتنی بر روش جستجوی شبکه‌ای**

محمود ی. شمس¹، احمد م. الشیوی²، السید م. الکناوی³، عبدالحمید ابراهیم⁴، فاطمه م. طلعت¹,⁵، زهراء طارق⁶

چکیده

کیفیت آب نقش حیاتی در زندگی انسان‌ها، حیوانات، گیاهان، صنایع و محیط‌زیست ایفا می‌کند. در دهه‌های اخیر، آلودگی و آلاینده‌ها تأثیر منفی بر کیفیت آب داشته‌اند. در این پژوهش، چالش اصلی پیش‌بینی شاخص کیفیت آب (WQI) و طبقه‌بندی کیفیت آب (WQC) است، به‌طوری که WQI یک شاخص مهم برای ارزیابی قابلیت استفاده از آب محسوب می‌شود.

در این مطالعه، بهینه‌سازی و تنظیم پارامترها برای بهبود دقت چندین مدل یادگیری ماشین مورد استفاده قرار گرفته است. روش‌های یادگیری ماشین به منظور پیش‌بینی WQI و WQC به کار گرفته شده‌اند. روش جستجوی شبکه‌ای (Grid Search) به‌عنوان یک روش اساسی برای بهینه‌سازی و تنظیم پارامترهای چهار مدل طبقه‌بندی و چهار مدل رگرسیون به کار رفته است.

در بخش مدل‌های طبقه‌بندی برای پیش‌بینی WQC، از چهار مدل زیر استفاده شده است:

جنگل تصادفی (RF)

تقویت گرادیان افزایشی (XGBoost)

تقویت گرادیان (GB)

تقویت تطبیقی (AdaBoost)

همچنین، در بخش مدل‌های رگرسیون برای پیش‌بینی WQI، از چهار مدل زیر بهره گرفته شده است:

رگرسیون نزدیک‌ترین همسایه (KNN Regressor)

رگرسیون درخت تصمیم (DT Regressor)

رگرسیون بردار پشتیبان (SVR)

رگرسیون پرسپترون چندلایه (MLP Regressor)

علاوه بر این، مرحله پیش‌پردازش داده‌ها شامل جایگزینی داده‌های گمشده (میانگین‌گیری) و نرمال‌سازی داده‌ها انجام شده است تا داده‌ها برای پردازش‌های بعدی مناسب‌سازی شوند. مجموعه داده‌های مورد استفاده در این پژوهش شامل 7 ویژگی و 1991 نمونه است.

برای ارزیابی کارایی مدل‌های طبقه‌بندی، پنج معیار ارزیابی محاسبه شده‌اند:

دقت (Accuracy)

بازخوانی (Recall)

دقت پیش‌بینی (Precision)

ضریب همبستگی متیوز (MCC)

امتیاز F1

همچنین، برای ارزیابی کارایی مدل‌های رگرسیون، چهار معیار محاسبه شده‌اند:

میانگین قدر مطلق خطا (MAE)

میانه قدر مطلق خطا (MedAE)

میانگین مربعات خطا (MSE)

ضریب تعیین (R²)

در بخش طبقه‌بندی، مدل GB با دقت 99.50٪ بهترین عملکرد را در پیش‌بینی مقادیر WQC ارائه داد. همچنین، بر اساس نتایج تجربی، مدل MLP Regressor در بخش رگرسیون عملکرد بهتری نسبت به سایر مدل‌ها داشت و توانست مقدار R² معادل 99.8٪ را در پیش‌بینی مقادیر WQI به دست آورد.

کلمات کلیدی

کیفیت آب · مدل‌های یادگیری ماشین · جستجوی شبکه‌ای · شاخص کیفیت آب · طبقه‌بندی کیفیت آب

# **.1 مقدمه**

آب یکی از ارزشمندترین منابعی است که تمام موجودات زنده به آن وابسته هستند. آلودگی آب منجر به کاهش کیفیت آن شده و سلامت موجودات دریایی و در نتیجه، سلامت انسان‌هایی که از آن‌ها تغذیه می‌کنند را تحت تأثیر قرار می‌دهد. این مسئله، نظارت بر کیفیت آب را برای حفظ حیات آبزیان و اکوسیستم‌های دریایی ضروری می‌سازد. درک مسائل و چالش‌های مرتبط با کیفیت آب برای کنترل و کاهش آلودگی آن نیز از اهمیت بالایی برخوردار است. در همین راستا، بسیاری از دولت‌ها در سراسر جهان برنامه‌هایی برای مدیریت اکولوژیکی آب تدوین کرده‌اند.

حدود یک میلیارد نفر در سراسر جهان به آب آشامیدنی سالم دسترسی ندارند و سالانه دو میلیون نفر به دلیل مصرف آب آلوده و شرایط نامناسب بهداشتی جان خود را از دست می‌دهند. بنابراین، حفظ کیفیت آب‌های شیرین یک اولویت حیاتی است. کیفیت نامناسب آب می‌تواند هزینه‌های سنگینی به همراه داشته باشد، زیرا منابع زیادی برای اصلاح زیرساخت‌های توزیع آب نیاز خواهد بود. ازاین‌رو، تقاضا برای مدیریت بهینه و کنترل کیفیت آب رو به افزایش است تا آب آشامیدنی سالم را با هزینه‌های مقرون‌به‌صرفه تأمین کند. به همین منظور، ارزیابی‌های سیستماتیک آب‌های شیرین، سیستم‌های دفع فاضلاب و روش‌های نظارتی سازمانی ضروری است.

اهمیت پیش‌بینی کیفیت آب

پیش‌بینی کیفیت آب به معنای برآورد تغییرات و نوسانات کیفیت آب در یک زمان مشخص است. ارزیابی کیفیت آب نقش مهمی در برنامه‌ریزی و تنظیم مقررات مرتبط با آب دارد. با پیش‌بینی تغییرات آتی در کیفیت آب، می‌توان استراتژی‌های مؤثرتری برای کنترل و جلوگیری از آلودگی آب تدوین کرد. همچنین، کیفیت آب باید در برنامه‌های انتقال و توزیع آب مورد ارزیابی قرار گیرد، زیرا مقادیر زیادی آب برای تأمین نیازهای روزانه مصرف می‌شود.

فناوری‌های هوش مصنوعی (AI) و یادگیری ماشین (ML) در حال حاضر نقش کلیدی در مسائل امنیتی ایفا می‌کنند و به جای استفاده از استراتژی‌های پیچیده، بر تحلیل روابط بین ورودی‌های سیستم و خروجی‌ها تمرکز دارند. پیش‌بینی کیفیت آب یکی از مؤلفه‌های اساسی برای برنامه‌ریزی، نظارت و کنترل آلودگی آب محسوب می‌شود. ازاین‌رو، توسعه یک روش کارآمد و عملی برای پیش‌بینی کیفیت آب امری ضروری است. همچنین، پیش‌بینی کیفیت آب در آینده می‌تواند از بروز تغییرات ناگهانی در کیفیت آب جلوگیری کرده و راه‌حل‌هایی برای مدیریت آن ارائه دهد. پیش‌بینی دقیق تغییرات کیفیت آب نه‌تنها تضمین‌کننده سلامت آب آشامیدنی انسان‌هاست، بلکه می‌تواند بهره‌وری شیلات را هدایت کرده و تنوع زیستی را حفظ کند.

نقش فناوری‌های نوین در نظارت بر کیفیت آب

ظهور فناوری‌هایی مانند سنجش از دور (RS)، رایانش ابری، اینترنت اشیا (IoT)، داده‌های کلان (Big Data) و هوش مصنوعی، فرصت‌های جدیدی برای بهبود و پیاده‌سازی روش‌های نظارت بر محیط‌زیست آبی ایجاد کرده است. در کشورهای مختلف، روش‌های تشخیص هوشمند برای حفظ محیط‌زیست آبی توسعه داده شده‌اند که شامل ایستگاه‌های پایش خودکار هیدرولوژیکی و کیفیت آب، شبکه‌های حسگر بی‌سیم (WSN)، سیستم‌های سنجش از دور، کشتی‌های نظارتی و ربات‌های زیرآبی پیشرفته می‌شود.

شاخص کیفیت آب (WQI) معیاری است که برای ارزیابی کیفیت آب در کاربردهای مختلف مورد استفاده قرار می‌گیرد. این شاخص می‌تواند تعیین کند که آیا آب برای مصرف انسانی، استفاده صنعتی یا زندگی آبزیان مناسب است یا خیر. هرچه مقدار WQI بالاتر باشد، کیفیت آب بهتر است. همچنین، طبقه‌بندی کیفیت آب (WQC) بر اساس دامنه مقادیر WQI توسعه یافته است که آب را در دسته‌های مختلفی از جمله آلوده یا تمیز قرار می‌دهد.

محاسبه WQI نیازمند ارزیابی چندین ویژگی فیزیکی، شیمیایی و بیولوژیکی آب مانند pH، اکسیژن محلول، کدورت، میزان مواد مغذی و وجود آلاینده‌ها است. با تجمیع این متغیرها، WQI یک ارزیابی جامع از کیفیت آب ارائه می‌دهد که در فرآیندهای تصمیم‌گیری مرتبط با مدیریت منابع آبی مؤثر است.

دلایل استفاده از یادگیری ماشین در پیش‌بینی کیفیت آب

روش‌های سنتی ارزیابی کیفیت آب معمولاً پر هزینه، زمان‌بر و فاقد قابلیت ارائه اطلاعات لحظه‌ای (Real-time) هستند. ازاین‌رو، مدل‌های پیش‌بینی مبتنی بر یادگیری ماشین، جایگزین مناسبی برای روش‌های سنتی محسوب می‌شوند. پیش‌بینی WQI و WQC به کمک مدل‌های یادگیری ماشین از چند جنبه مهم است:

* پایش لحظه‌ای کیفیت آب: مدل‌های پیش‌بینی امکان برآورد WQI و WQC را به‌صورت هم‌زمان یا نزدیک به لحظه واقعی فراهم می‌کنند که کارآمدتر و مقرون‌به‌صرفه‌تر از روش‌های آزمایشگاهی سنتی است.
* مدیریت داده‌های ناقص: برخی داده‌های مربوط به کیفیت آب ممکن است ناقص یا گم‌شده باشند. مدل‌های پیش‌بینی می‌توانند با استفاده از داده‌های موجود، مقادیر از دست‌رفته را تخمین بزنند و از این طریق، دقت محاسبات WQI را افزایش دهند.
* بهینه‌سازی منابع: با پیش‌بینی دقیق‌تر WQI و WQC، می‌توان منابع را بهینه‌تر تخصیص داد. تصمیم‌گیرندگان می‌توانند تلاش‌های نمونه‌گیری را اولویت‌بندی کرده، نظارت را به مناطق مهم هدایت کنند و راهبردهای تصفیه آب را بهینه‌سازی نمایند.
* سیستم‌های هشدار زودهنگام: مدل‌های پیش‌بینی می‌توانند به‌عنوان پایه‌ای برای توسعه سیستم‌های هشدار زودهنگام برای مشکلات کیفیت آب عمل کنند. با پایش و پیش‌بینی مداوم شاخص‌های کیفیت آب، می‌توان خطرات بالقوه را پیش از وقوع شناسایی کرده و اقدامات پیشگیرانه‌ای برای کاهش اثرات آلودگی انجام داد.

روش‌های مورد استفاده در این پژوهش

در این پژوهش، از الگوریتم‌های یادگیری ماشین برای پیش‌بینی شاخص کیفیت آب (WQI) و طبقه‌بندی کیفیت آب (WQC) استفاده شده است. روش جستجوی شبکه‌ای (Grid Search) به عنوان یک روش مهم برای بهینه‌سازی و تنظیم پارامترهای مدل‌ها مورد استفاده قرار گرفته است.

مدل‌های طبقه‌بندی برای پیش‌بینی WQC شامل موارد زیر هستند:

* جنگل تصادفی (RF)
* تقویت گرادیان افزایشی (XGBoost)
* تقویت گرادیان (GB)
* تقویت تطبیقی (AdaBoost)

مدل‌های رگرسیون برای پیش‌بینی WQI شامل موارد زیر هستند:

* رگرسیون نزدیک‌ترین همسایه (KNN Regressor)
* رگرسیون درخت تصمیم (DT Regressor)
* رگرسیون بردار پشتیبان (SVR)
* رگرسیون پرسپترون چندلایه (MLP Regressor)

نتایج آزمایش‌ها نشان داد که مدل GB در بخش طبقه‌بندی با دقت 99.5٪ و مدل MLP Regressor در بخش رگرسیون با مقدار R² برابر 99.8٪، بهترین عملکرد را داشتند.

ساختار مقاله

باقی‌مانده مقاله به‌صورت زیر سازماندهی شده است:

* بخش 2: بررسی مطالعات مرتبط با پیش‌بینی کیفیت آب
* بخش 3: مواد و روش‌های پیشنهادی
* بخش 4: تشریح روش‌شناسی تحقیق
* بخش 5: ارائه نتایج و بحث
* بخش 6: نتیجه‌گیری

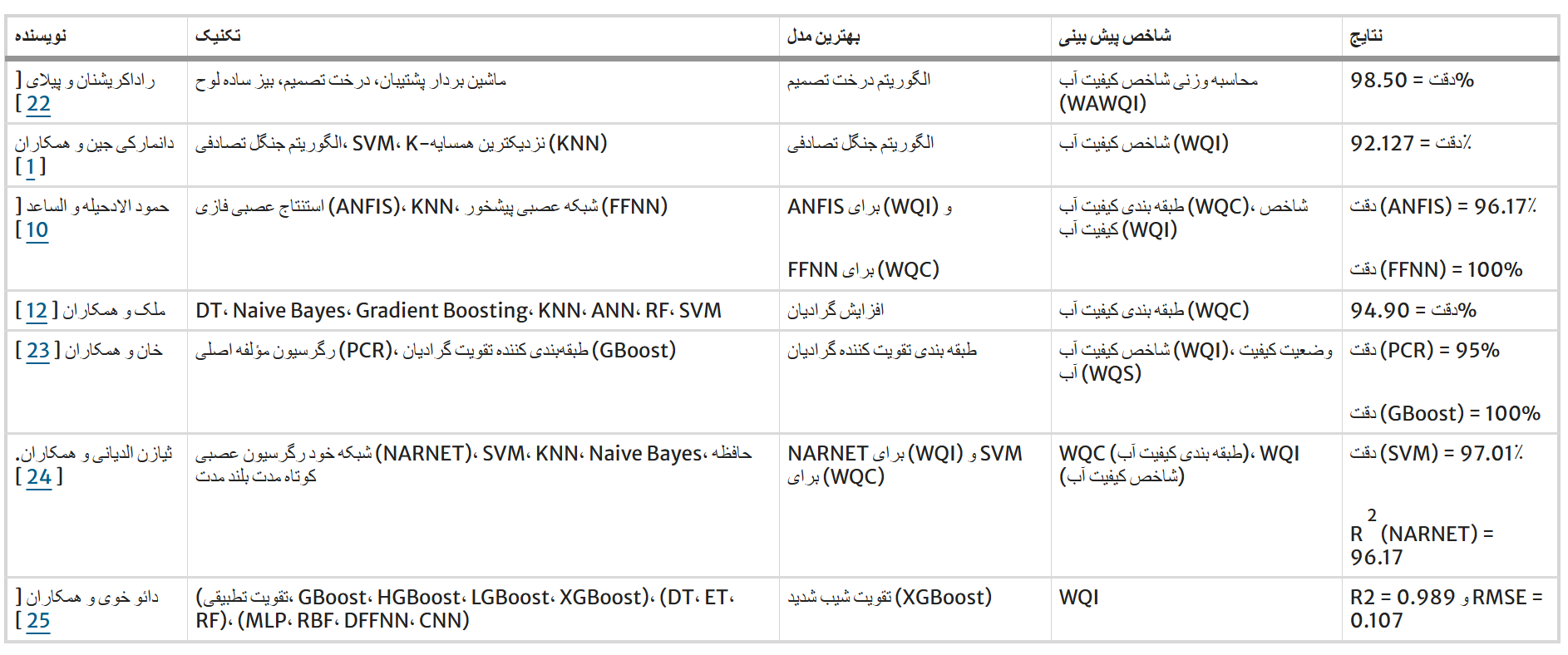
**۲. کارهای مرتبط**

شبکه‌های عصبی مصنوعی (ANN)، رگرسیون‌های برداری پشتیبان (SVR)، سیستم‌های خاکستری (GS)، تحلیل‌های رگرسیونی (RA) و سایر روش‌ها به طور معمول برای برآورد کیفیت آب استفاده می‌شوند [3]. لیو و همکاران [9] کیفیت آب آشامیدنی حوضه رودخانه یانگ‌تسه را با استفاده از شبکه حافظه بلندمدت-کوتاه‌مدت (LSTM) پیش‌بینی کردند. اکسیژن حل‌شده (DO)، pH، نیاز شیمیایی به اکسیژن (COD) و NH3-N برای ساخت الگوریتم LSTM استفاده شدند. تکنیک LSTM قابلیت‌هایی برای نظارت بر کیفیت آب نشان داده است.

ساکشی خولار و ناهنی سینگ [15] یک مدل Bi-LSTM مبتنی بر یادگیری عمیق (DLBL-WQA) را برای پیش‌بینی متغیرهای کیفیت آب رودخانه یامونا در هند ارائه دادند. یک مقایسه نشان داد که رویکرد پیشنهادی در مقایسه با سایر روش‌ها از نظر نرخ خطا و دقت پیش‌بینی برتری دارد. ثانی عبای و همکاران [16] چهار تکنیک یادگیری ماشین شامل استنتاج عصبی-فازی (ANFIS)، پس‌انتشار (BPNN)، پرسپترون چند لایه (MLP) و رگرسیون برداری پشتیبان (SVR) را برای پیش‌بینی شاخص کیفیت آب (WQI) بررسی کردند. نتایج به دست آمده نشان داد که تکنیک‌های هوشمند ساخته‌شده برای پیش‌بینی WQI در سه ایستگاه با استفاده از نتایج مدل‌سازی بهتر از ترکیب شبکه‌های عصبی (NNE) موفق بوده‌اند و بنابراین می‌توانند به عنوان یک استراتژی پیش‌بینی قابل اعتماد استفاده شوند.

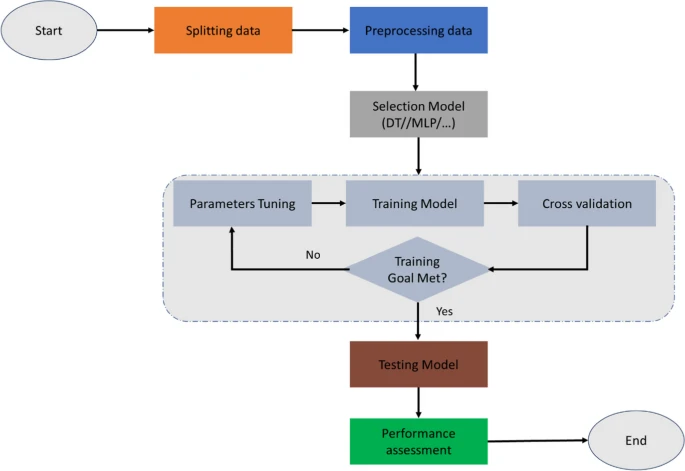
البته، البلتاجی و همکاران [17] از چهار تکنیک مستقل شامل مدل درخت M5P (M5P)، رگرسیون افزایشی (AR)، ماشین بردار پشتیبان (SVM) و فضای تصادفی (RSS) برای پیش‌بینی WQI با استفاده از استراتژی حذف متغیرها استفاده کردند. رگرسیون افزایشی (AR) از سایر روش‌های داده‌محور برتری داشت. AR به عنوان یک رویکرد بهینه با نتایج خوب به دلیل قابلیت پیش‌بینی بهتر با کمترین تعداد متغیرهای منبع معرفی شد و بنابراین می‌تواند برای پیش‌بینی قابل اعتماد و دقیق WQI در حوضه آکوت استفاده شود. سید اسدالله و همکاران [18] رگرسیون درخت اضافی (ETR)، یک تکنیک یادگیری ماشین ترکیبی، را برای پیش‌بینی نرخ‌های ماهانه WQI در امتداد رودخانه لام تسوئن در هنگ‌کنگ ارائه دادند. مقایسه نتایج بین ETR و روش‌های مستقل معمولی (SVR، DTR) نشان داد که رویکرد ETR پیش‌بینی‌های WQI قابل اعتمادتری در مراحل آموزش و آزمایش ارائه می‌دهد. به طور کلی، روش ETR از تکنیک‌های پیشین برای پیش‌بینی WQI از نظر دقت پیش‌بینی و تعداد متغیرهای ورودی پیشی گرفت. علاوه بر این، نوسیر 2022 و همکاران [19] یک مدل رگرسیونی پیش‌بینی مبتنی بر استراتژی اصلی با استفاده از شاخص‌های SWI و روش‌های هوش مصنوعی (AI) برای نظارت بر شور شدن آب‌های زیرزمینی ناشی از نفوذ آب شور (SWI) در سفره آب‌های زیرزمینی دلتای شرقی نیل، مصر ارائه دادند. فرید گاراباغی و همکاران [20] چهار تکنیک یادگیری ماشین با رویکردهای یادگیری ترکیبی شامل جنگل تصادفی، LogitBoost، XGBoost و AdaBoost را برای طبقه‌بندی کیفیت آب معرفی کردند. در نتیجه، XGBoost از سایر روش‌های طبقه‌بندی پیشی گرفت و با دقت 96.9696 درصد زمانی که ویژگی‌های مهم در مرحله طبقه‌بندی گنجانده شدند، برتری داشت. مدل XGBoost به عنوان بهترین روش طبقه‌بندی با دقت بالا و 95.606 درصد با اعتبارسنجی متقاطع ده‌برابری پیشنهاد می‌شود زمانی که مرحله طبقه‌بندی شامل هفت متغیر انتخاب‌شده با استفاده از انتخاب ویژگی حذف به عقب باشد. مهدی حسن و همکاران [21] الگوریتم‌های یادگیری ماشین مانند NN، RF، SVM، BTM و MLR را برای طبقه‌بندی مجموعه داده‌های کیفیت آب در مکان‌های مختلف در سراسر هند به کار بردند. نیاز به اکسیژن زیستی (BOD)، اکسیژن حل‌شده (DO)، کلریفیکاتورهای کلی (TC)، pH، نیترات و هدایت الکتریکی (EC) از جمله عواملی هستند که بر کیفیت آب تأثیر می‌گذارند. این ویژگی‌ها در 5 مرحله پردازش می‌شوند: نرمال‌سازی مین-ماکس برای پیش‌پردازش داده‌ها و نگهداری داده‌های مفقود با استفاده از RF، همبستگی ویژگی‌ها، طبقه‌بندی یادگیری ماشین به کار رفته و اهمیت طبقه‌بندی ویژگی‌ها. حداکثر دقت این مطالعه، دقت بالاتر، کاپا و دقت پایین به ترتیب 99.83، 99.99، 99.17 و 99.07 بود. نتایج نشان داد که هدایت الکتریکی، نیترات، DO، PH، BOD و TC ویژگی‌های اصلی هستند که به طبقه‌بندی کیفیت آب کمک می‌کنند و نتایج اهمیت ویژگی‌ها به ترتیب 81.494، 74.78، 105.770، 36.805، 130.173 و 105.166 بودند. جدول 1 برخی از مدل‌های یادگیری ماشین برای پیش‌بینی کیفیت آب را فهرست می‌کند.

طبق کارهای قبلی، دقت پیش‌بینی و طبقه‌بندی با استفاده از تکنیک‌های یادگیری ماشین بهبود یافته است، بنابراین در بخش بعدی به تأثیر برخی از تکنیک‌های یادگیری ماشین برای پیش‌بینی کیفیت آب با دقت بالا در پیش‌بینی و طبقه‌بندی پرداخته می‌شود.



۳. مواد و روش‌ها

پس از پیش‌پردازش داده‌های اولیه، یک رویکرد خاص یادگیری ماشین (ML) برای آموزش و تأیید با استفاده از مجموعه‌های آموزش و اعتبارسنجی انتخاب می‌شود. قبل از آزمایش، متغیرهای هایپر مربوطه به‌طور دقیق تنظیم خواهند شد تا زمانی که هدف از پیش تعیین‌شده آموزش برآورده شود. در نهایت، مجموعه داده‌های آزمایشی برای ارزیابی رویکرد آموزش‌دیده و بررسی بهبود آن اعمال خواهد شد. برای شفافیت، نمودار جریان مدل‌سازی ML در شکل 1 ارائه شده است. نمودار بلوکی کلی مدل‌های ML با تقسیم‌بندی داده‌ها و پیش‌پردازش آغاز می‌شود، سپس انتخاب مدل انجام می‌شود. مدل انتخاب‌شده سپس تحت آموزش، آزمایش و اعتبارسنجی قرار می‌گیرد. اعتبارسنجی متقابل برای ارزیابی اینکه آیا مدل آموزشی به اهداف خود رسیده است یا خیر، استفاده می‌شود. اگر چنین باشد، مدل می‌تواند به مرحله آزمایش و ارزیابی عملکرد ادامه دهد. در غیر این صورت، پارامترهای مدل نیاز به تنظیم دقیق بیشتر در طول آموزش دارند. برای افزایش اثربخشی پیش‌بینی کیفیت آب در این کار، هشت رویکرد پرکاربرد ML اصلاح‌شده، پیاده‌سازی‌شده و استفاده شده‌اند که در ادامه آورده شده است.

شکل. 1

فلوچارت کلی مدل های یادیگری ماشین

## **۳.۱ مدل دسته‌بندی برای پیش‌بینی WQC**

در این بخش چهار الگوریتم دسته‌بندی معرفی شده‌اند: RF، XGBoost، GB و AdaBoost.

## **۳.۱.۱ جنگل تصادفی (RF)**

روش RF یک تکنیک گروهی است که برای دسته‌بندی استفاده می‌شود. این یک روش یادگیری ماشین نظارت‌شده است که از درختان تصمیم متعدد تشکیل شده است. چون این یک تکنیک گروهی است، از بهترین نتیجه‌ای که توسط درختان تصمیم مختلف ارائه می‌شود استفاده می‌کند، که اشتباهات عمومی‌سازی را کاهش می‌دهد و محدود می‌کند زیرا حجم معماری درخت در جنگل افزایش می‌یابد [26]. الگوریتم درخت دسته‌بندی و رگرسیون (CART) توسط درخت تصمیم برای دسته‌بندی مجموعه‌های داده بر اساس پارامتر هدف استفاده می‌شود. این روش در کنار تکنیک bagging برای اهداف بازنمونه‌گیری استفاده می‌شود، که داده‌های آموزشی را به‌هنگام تشکیل درخت جدید به‌روزرسانی می‌کند [27].

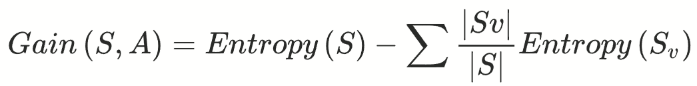
براساس پارامترها و معادلات زیر، یک ساختار درختی برای دسته‌بندی ویژگی‌ها ساخته می‌شود [1]. شاخص جینی می‌تواند برای ساخت درخت تصمیم برای هر مجموعه S استفاده شود و با فرمول زیر محاسبه می‌شود:

(1)

A black and white symbol

AI-generated content may be incorrect.همچنین آنتروپی و بهره‌برداری اطلاعات هنگام ساخت درخت تصمیم و تعیین نتیجه آن اهمیت دارند. این مقادیر می‌توانند با استفاده از فرمول‌های زیر محاسبه شوند:

(2)

که در آن p کسری از S است که به کلاس ‘i’ تعلق دارد، برای هر مجموعه S داده‌شده.

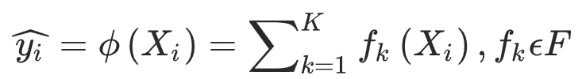
(3)

که در آن Sv زیرمجموعه‌ای از S است که در آن پارامتر A دارای مقدار v است.

RF مزایای متعددی ارائه می‌دهد. این روش از مشکل هم‌خطی چندمتغیره که از معایب تحلیل رگرسیون معمولی است جلوگیری می‌کند. همچنین در رگرسیون و دسته‌بندی عملکرد عالی دارد و تسلط قوی بر داده‌های چندبعدی دارد [28].

## **۳.۱.۲ تقویت گرادیان افراطی (XGBoost)**

XGBoost یک روش تقویت درخت تصمیم است که با روش کلاسیک تقویت درخت تصمیم گرادیان (GBDT) متفاوت است [29]. بر اساس مسأله بهینه‌سازی، GBDT استاندارد تنها از اطلاعات مشتق اول استفاده می‌کند. سپس تابع ضرر تحت گسترش دوم تیلور قرار می‌گیرد که از مشتقات اول و دوم استفاده می‌کند. تابع ضرر شامل یک عبارت تنظیمی است که پیچیدگی تکنیک را مدیریت می‌کند و از تطبیق بیش از حد (Overfitting) جلوگیری می‌کند. روش XGBoost به‌صورت زیر مشتق می‌شود [28]:



(4)



که در آن نشان‌دهنده فضای تابعی است که درخت تصمیم را تعریف می‌کند و تعداد گره‌های برگ درخت تصمیم است. تابع ضرر به‌صورت زیر است:



(5)

A black symbols on a white background

AI-generated content may be incorrect.

(6)

جزء اول در معادله (۵) تعداد برگ‌ها را نشان می‌دهد، در حالی که جزء دوم اندازه نتیجه را نشان می‌دهد. XGBoost برای هر گره درخت، Gain را محاسبه می‌کند تا ارزیابی کند که آیا شاخه تولید شده مرتبط است یا خیر.

A black text with a plus and a white background

AI-generated content may be incorrect.

(7)

که در آن نشان‌دهنده میزان واقعی Gain قبل از تقسیم و تعداد برگ‌های جدید است.

## **۳.۱.۳ مدل تقویت گرادیان (GB)**

GB یک روش یادگیری ماشین است که چندین روش طبقه‌بندی ضعیف، معمولاً درخت‌های تصمیم، را ترکیب می‌کند تا یک طبقه‌بند قابل اعتماد برای وظایف طبقه‌بندی و رگرسیون ایجاد کند. این سیستم را به‌صورت مرحله‌ای می‌سازد، مشابه سایر روش‌های تقویت، و آن را با حداکثر کردن یک تابع هزینه مناسب تعمیم می‌دهد. در روش GB، نمونه‌هایی که در یک مرحله به‌درستی شناسایی نشده‌اند، در مرحله بعدی وزن بیشتری دریافت می‌کنند. مزایای GB شامل دقت پیش‌بینی عالی و فرآیند سریع است [30]. این روش شباهت زیادی به روش تقویت تطبیقی (AdaBoost) دارد، اگرچه AdaBoost این ضعف را دارد که تحت تأثیر مقادیر پرت (outliers) قرار می‌گیرد و به راحتی توسط داده‌های نویزی غالب می‌شود [31].

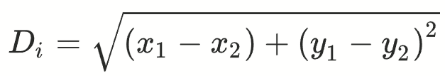
## **۳.۱.۴ مدل تقویت تطبیقی (AdaBoost)**

روش AdaBoost با ترکیب چندین یادگیرنده ضعیف در یک یادگیرنده قوی، عملکرد طبقه‌بند را بهبود می‌بخشد. این روش به‌طور مکرر وزن‌های نمونه‌ها را بر اساس اشتباهات طبقه‌بندی تنظیم می‌کند، وزن نمونه‌های اشتباه طبقه‌بندی‌شده را افزایش داده و وزن نمونه‌های درست طبقه‌بندی‌شده را کاهش می‌دهد. در نتیجه، روش‌های طبقه‌بندی که بر داده‌های اشتباه طبقه‌بندی‌شده تمرکز دارند، به جای نمونه‌های کلاس اقلیت، استفاده می‌شوند. از آنجا که AdaBoost بر عملکرد پیش‌بینی تمرکز دارد، این روش به‌سمت کلاس اکثریت گرایش دارد که به عملکرد پیش‌بینی کلی بیشتر کمک می‌کند [32].

## **۳.۲ مدل‌های رگرسیون برای پیش‌بینی شاخص کیفیت آب (WQI)**

در این بخش، چهار الگوریتم رگرسیون شامل KNN، درخت تصمیم (DT)، رگرسیون بردار پشتیبان (SVR) و پرسپترون چندلایه (MLP) معرفی شده‌اند.

## **۳.۲.۱ مدل نزدیک‌ترین همسایگان (KNN)**

تکنیک KNN نمونه‌ها را با یافتن نزدیک‌ترین نقاط همسایه و اختصاص اکثریت n همسایه به یک کلاس متمایز می‌کند. در صورت تساوی، روش‌های مختلفی ممکن است برای حل آن به کار گرفته شود. با این حال، KNN برای مجموعه داده‌های بزرگ توصیه نمی‌شود، زیرا تمامی محاسبات را در طول تست انجام می‌دهد و در تمام داده‌های آموزش‌دیده‌شده همگرا می‌شود و در هر بار نزدیک‌ترین همسایه را محاسبه می‌کند [33]. برای یافتن نزدیک‌ترین همسایه در بردار ویژگی‌ها، از تابع فاصله اقلیدسی (Di) به صورت زیر استفاده شده است:

(8)

که در آن  
نشان‌دهنده پارامترهای ورودی داده است.

## **۳.۲.۲ درخت تصمیم (DT)**

درخت تصمیم (DT) یک روش ساده و پایه‌ای است که بر اساس مقادیر تمام متغیرهای ورودی مرتبط، تصمیم‌گیری می‌کند. DT پارامتر ریشه را بر اساس آنتروپی انتخاب کرده و سپس وزن سایر متغیرها را تجزیه و تحلیل می‌کند. این مدل تمامی تصمیمات متغیرها را در یک ساختار درختی از بالا به پایین گروه‌بندی کرده و تصمیم‌گیری را بر اساس مقادیر مختلف ویژگی‌های خاص انجام می‌دهد. تحقیقات پیشین نشان داده‌اند که مدل‌های درخت تصمیم در داده‌های نامتوازن عملکرد خوبی دارند. با این حال، تکنیک‌های مبتنی بر مجموعه درخت‌های تصمیم، مانند تقویت گرادیان (GB) و جنگل تصادفی (RF)، تقریباً همیشه عملکرد بهتری نسبت به یک درخت تصمیم منفرد دارند [12].

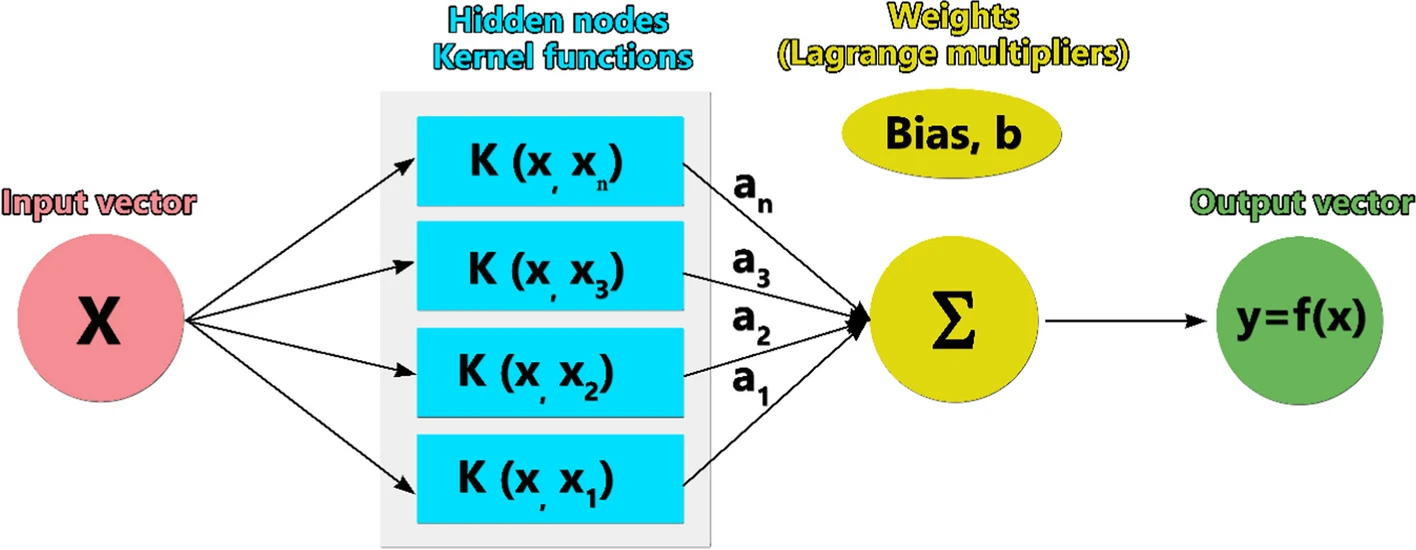
مزایای مدل‌های مبتنی بر درخت تصمیم شامل عدم حساسیت به مقادیر از دست رفته، توانایی حفظ ویژگی‌ها و داده‌های اصلی، و کارایی بالا است. در مقایسه با سایر الگوریتم‌های یادگیری ماشین، تکنیک‌های مبتنی بر درخت تصمیم برای پیش‌بینی کوتاه‌مدت مناسب‌تر هستند و ممکن است سرعت محاسباتی بالاتری داشته باشند [34].

## **۳.۲.۳ رگرسیون بردار پشتیبان (SVR)**

رگرسیون بردار پشتیبان (SVR) یک تکنیک یادگیری ماشین است که از ماشین بردار پشتیبان (SVM) مشتق شده و به عنوان روشی امیدوارکننده برای حل مسائل غیرخطی مانند رگرسیون، پیش‌بینی، طبقه‌بندی و برآورد تابع در نظر گرفته می‌شود. این تکنیک یک روش کارآمد برای حل مسائل برنامه‌نویسی درجه دوم محدب است. علاوه بر این، SVR دارای ویژگی‌های برجسته‌ای مانند عدم همگرایی به بهینه محلی، فرمول‌بندی ریاضی قوی، قابلیت پیش‌بینی بالا و مقیاس‌پذیری مناسب است. با این حال، مجموعه داده‌های آموزشی باید به صورت دستی برچسب‌گذاری شوند و سه متغیر تکنیک SVR باید با استفاده از اطلاعات قبلی تنظیم شوند [35،36،37]. تابع کلی غیرخطی در SVR به صورت زیر تعریف می‌شود:

(9)

که y در آن نشان‌دهنده رابطه بین متغیر پیش‌بینی‌شده و متغیرهای پیش‌بینی‌کننده، بردار وزن، تابع نگاشت غیرخطی مجموعه داده‌های ورودی، و مقدار آستانه اسکالر را نشان می‌دهد. شکل ۲ ساختار SVR را نمایش می‌دهد.

شکل. 2

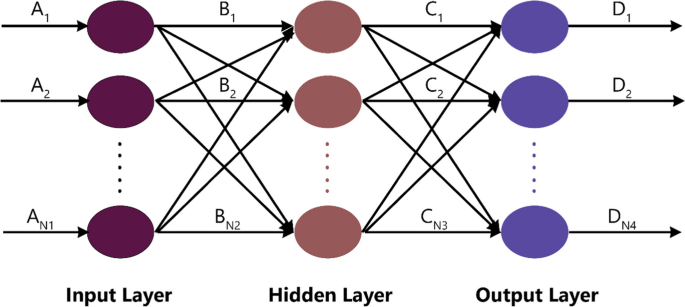
ساختمان مدل SVR

## **۳.۲.۴ رگرسور پرسپترون چندلایه (MLP)**

پرسپترون چندلایه (MLP) شامل یک لایه ورودی، یک لایه خروجی و چندین لایه پنهان است. سیگنال ورودی از طریق لایه ورودی به لایه پنهان منتقل شده و در آنجا، نورون‌ها از نظر محاسباتی پردازش می‌شوند. سپس، خروجی این لایه به سمت لایه خروجی هدایت می‌شود. خروجی شبکه عصبی MLP فقط به ورودی فعلی وابسته است و تأثیری از ورودی‌های قبلی یا آینده نمی‌پذیرد؛ در نتیجه، این شبکه عصبی به عنوان یک شبکه عصبی پیش‌خور چندلایه نیز شناخته می‌شود.

شبکه‌های عصبی MLP یکی از طرح‌های متداول شبکه‌های عصبی محسوب می‌شوند که دارای ساختار ساده، اجرای آسان، تحمل خطای قوی، انعطاف‌پذیری بالا، مقیاس‌پذیری مناسب و قابلیت برجسته در نگاشت غیرخطی هستند [7].شکل ۳ معماری شبکه عصبی MLP را نمایش می‌دهد.

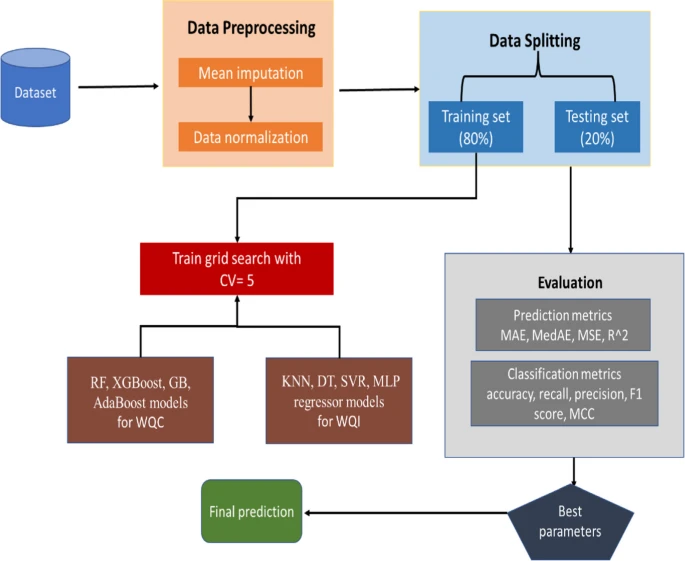
شکل. 3



توپولوژی شبکه های عصبی MLP

## **۴. روش پیشنهادی**

آلودگی آب یکی از جدی‌ترین مسائل زیست‌محیطی پیش روی بشریت است و خسارات ناشی از آن عمدتاً به عدم پیش‌بینی، هشدار زودهنگام و مدیریت اضطراری بازمی‌گردد. بنابراین، پیاده‌سازی یک سیستم نظارتی و هشدار زودهنگام مناسب جهت اتخاذ تصمیمات هوشمندانه و مدیریت کیفیت آب، یک چالش علمی و فنی اساسی محسوب می‌شود که باید به سرعت مورد توجه قرار گیرد [38]. در سال‌های اخیر، روش‌های یادگیری ماشین به‌سرعت پیشرفت کرده‌اند. شکل ۴ روش پیشنهادی برای پیش‌بینی کیفیت آب را نمایش می‌دهد.

شکل. 4

روش شناسایی پیشنهادی

روش پیشنهادی در این پژوهش، یک مدل یادگیری ماشین برای ارزیابی کیفیت آب توسعه داده است که بر یک مجموعه‌ داده شامل هفت ویژگی زیر استوار است:

* اکسیژن محلول
* pH
* هدایت الکتریکی
* اکسیژن مورد نیاز بیولوژیکی (BOD)
* نیترات
* کلیفرم مدفوعی
* کلیفرم کل

این مجموعه داده پیش‌پردازش شده است که شامل جایگزینی مقادیر گمشده با میانگین و نرمال‌سازی داده‌ها می‌شود. داده‌ها به مجموعه‌های آموزشی (۸۰٪) و آزمایشی (۲۰٪) تقسیم شده‌اند. در مرحله آموزش، جستجوی شبکه‌ای (Grid Search) همراه با اعتبارسنجی متقاطع (CV=5) برای تنظیم ابرپارامترهای چهار مدل طبقه‌بندی (RF، XGBoost، GB، و AdaBoost) و چهار مدل رگرسیون (KNN، DT، SVR، و MLP) به‌کار گرفته شده است.

انتخاب مدل‌های یادگیری ماشین

انتخاب مدل‌های خاص برای روش ارزیابی کیفیت آب به ویژگی‌های داده، مسئله مورد بررسی و نیازمندی‌های عملکردی مدل بستگی دارد. مدل‌های منتخب برای این پژوهش به دلیل توانایی آن‌ها در پردازش ویژگی‌های مجموعه‌ داده‌های کیفیت آب و عملکرد مطلوب آن‌ها در مسائل مشابه انتخاب شده‌اند. مدل‌های یادگیری گروهی (Ensemble) که در این مطالعه به‌کار رفته‌اند، با ترکیب چندین یادگیرنده ضعیف، یک مدل قوی‌تر را ایجاد می‌کنند. این مدل‌ها اغلب در مسائل طبقه‌بندی با ویژگی‌های زیاد و تعاملات پیچیده بین متغیرهای ورودی و خروجی مورد استفاده قرار می‌گیرند. روش‌های گروهی می‌توانند این تعاملات پیچیده را ثبت کرده و دقت مدل را افزایش دهند.

* جنگل تصادفی (RF) به دلیل توانایی آن در مدیریت داده‌های با ابعاد بالا و جلوگیری از بیش‌برازش (Overfitting) شناخته شده است.
* XGBoost، GB، و AdaBoost به دلیل سرعت بالای آموزش و پیش‌بینی و همچنین دقت بالا در بسیاری از مسائل مورد استفاده قرار می‌گیرند.

برای مدل‌های رگرسیونی، چهار الگوریتم KNN، DT، SVR، و MLP انتخاب شده‌اند:

* KNN یک مدل غیرپارامتریک است که توانایی پردازش رابطه‌های خطی و غیرخطی بین متغیرهای ورودی و خروجی را دارد.
* درخت تصمیم (DT) یک مدل مبتنی بر درخت است که می‌تواند ارتباطات غیرخطی را مدیریت کند و تفسیر آن ساده است.
* رگرسیون بردار پشتیبان (SVR) یک مدل مبتنی بر هسته (Kernel-Based) است که در مجموعه داده‌های کوچک عملکرد خوبی دارد و قابلیت مدیریت ارتباطات غیرخطی را داراست.
* پرسپترون چندلایه (MLP) یک مدل مبتنی بر شبکه عصبی است که می‌تواند تعاملات پیچیده بین متغیرهای ورودی و متغیر هدف را مدیریت کند.

ارزیابی عملکرد مدل‌ها

در مرحله آزمایش، عملکرد مدل‌ها با استفاده از معیارهای مختلف سنجش خطا و دقت ارزیابی شده است:

* برای مدل‌های رگرسیونی:
  + میانگین قدرمطلق خطا (MAE)
  + میانه قدرمطلق خطا (MedAE)
  + میانگین مربعات خطا (MSE)
  + ضریب تعیین (R²)
* برای مدل‌های طبقه‌بندی:
  + دقت (Accuracy)
  + بازخوانی (Recall)
  + دقت پیش‌بینی (Precision)
  + امتیاز F1
  + ضریب همبستگی متیوز (MCC)

تنظیم ابرپارامترها با جستجوی شبکه‌ای (Grid Search)

جستجوی شبکه‌ای (Grid Search) یک روش تنظیم ابرپارامترها در یادگیری ماشین است که به منظور یافتن بهترین ترکیب از ابرپارامترها برای یک مدل خاص مورد استفاده قرار می‌گیرد. ابرپارامترها مقادیری هستند که قبل از آموزش مدل باید تعیین شوند و از داده‌ها قابل یادگیری نیستند.

نمونه‌هایی از ابرپارامترها شامل موارد زیر هستند:

* نرخ یادگیری (Learning Rate)
* پارامتر تنظیم (Regularization Parameter)
* تعداد لایه‌های یک شبکه عصبی
* تعداد درخت‌ها در یک جنگل تصادفی

مراحل جستجوی شبکه‌ای به شرح زیر است:

1. تعریف ابرپارامترها و مقادیر یا محدوده‌های ممکن آن‌ها
2. ایجاد یک شبکه از تمام ترکیب‌های ممکن ابرپارامترها
3. برای هر ترکیب از ابرپارامترها در شبکه:
   * آموزش مدل روی مجموعه داده‌های آموزشی
   * ارزیابی مدل با استفاده از مجموعه اعتبارسنجی متقاطع (CV=5)
   * ثبت عملکرد مدل بر اساس یک معیار مشخص
4. انتخاب ترکیب ابرپارامتری که بهترین عملکرد را داشته است

جستجوی شبکه‌ای از نظر محاسباتی ممکن است هزینه‌بر باشد، خصوصاً زمانی که تعداد ابرپارامترها و محدوده‌های ممکن آن‌ها زیاد باشد. برای کاهش هزینه‌های محاسباتی، می‌توان از روش جستجوی تصادفی (Randomized Search) استفاده کرد که در آن یک زیرمجموعه تصادفی از ابرپارامترها نمونه‌گیری شده و بررسی می‌شود.

## **۴.۱ مجموعه داده**

مجموعه داده مورد استفاده در این پژوهش در لینک زیر در دسترس است:  
<https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data>

این مجموعه داده از دریاچه‌ها و رودخانه‌های هند، در بازه زمانی ۲۰۰۵ تا ۲۰۱۴، در چندین موقعیت جغرافیایی جمع‌آوری شده است. دولت هند این داده‌ها را با هدف اطمینان از سلامت و قابلیت شرب آب گردآوری کرده است. مجموعه داده شامل ۱۹۹۱ نمونه و ۷ ویژگی است.

ویژگی‌های این مجموعه داده عبارتند از:

* اکسیژن محلول (Dissolved Oxygen): نشان‌دهنده سطح اکسیژن حل‌شده در آب است که برای حفظ حیات آبزیان ضروری است.
* pH: سطح اسیدی یا بازی بودن آب را مشخص می‌کند که معیاری برای تعادل شیمیایی آب محسوب می‌شود.
* هدایت الکتریکی (Conductivity): ظرفیت آب در هدایت جریان الکتریکی را ارزیابی کرده و اطلاعاتی درباره وجود مواد جامد محلول در آب ارائه می‌دهد.
* اکسیژن مورد نیاز بیولوژیکی (Biological Oxygen Demand - BOD): میزان اکسیژن محلول مصرف‌شده توسط میکروارگانیسم‌ها در آب را اندازه‌گیری کرده و نشان‌دهنده میزان آلودگی آلی آب است.
* نیترات (Nitrate): غلظت یون نیترات در آب را بررسی می‌کند که می‌تواند نشانگر آلودگی ناشی از کودهای شیمیایی یا فاضلاب باشد.
* کلیفرم مدفوعی (Fecal Coliform): نشانه‌ای از آلودگی مدفوعی است، زیرا وجود باکتری‌های کلیفرم در آب را منعکس می‌کند.
* کلیفرم کل (Total Coliform): نشان‌دهنده میزان کلی باکتری‌های کلیفرم، اعم از منشأ مدفوعی و غیرمدفوعی در آب است.

پیش‌پردازش داده‌ها

به‌منظور اطمینان از کیفیت و قابلیت استفاده از مجموعه داده در این پژوهش، برخی فرآیندهای پیش‌پردازش روی داده‌ها اعمال شده است. این فرآیندها شامل مدیریت داده‌های گمشده و حذف داده‌های پرت است که از مشکلات مهم در مجموعه داده‌های واقعی محسوب می‌شوند. با این حال، جزئیات دقیق مراحل پیش‌پردازش داده‌ها در متن ارائه نشده است.

تحلیل آماری داده‌ها

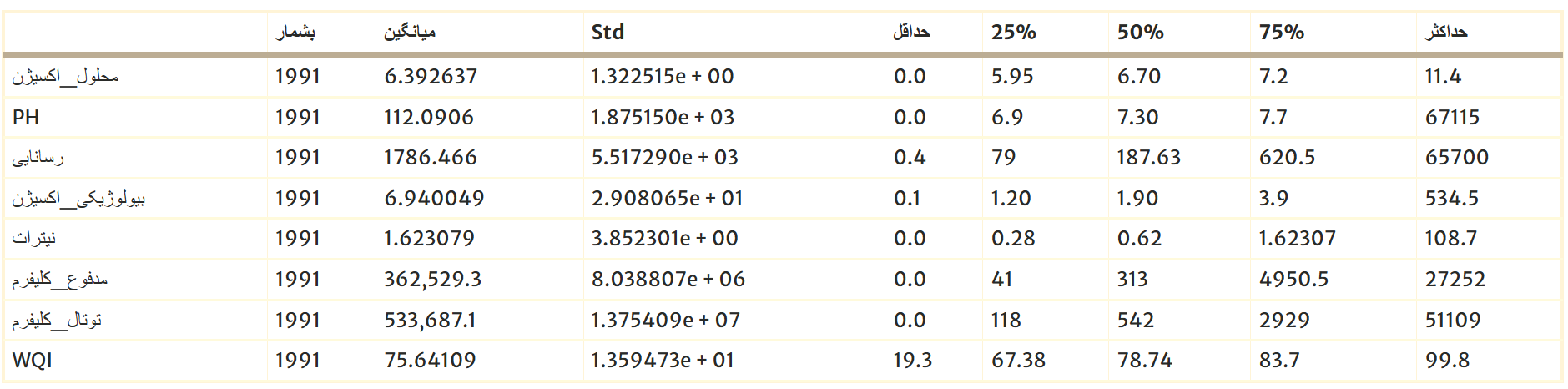
علاوه بر این، همان‌طور که در جدول ۲ نمایش داده شده است، در این پژوهش محاسبات آماری روی ویژگی‌های مجموعه داده انجام شده است. این محاسبات شامل معیارهایی مانند:

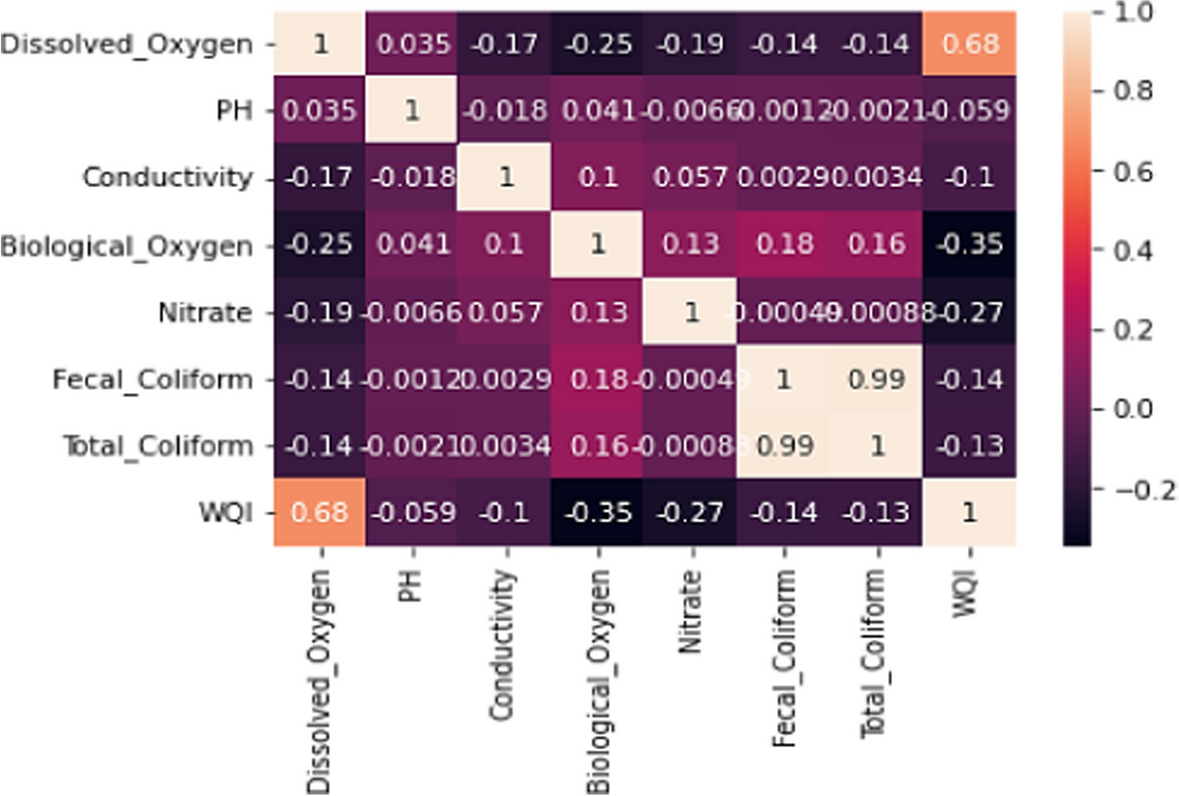
* میانگین (Mean)
* انحراف معیار (Standard Deviation)
* حداقل و حداکثر مقدار (Minimum & Maximum)
* چارک‌ها (Quartiles)

این معیارها اطلاعات ارزشمندی درباره توزیع و خصوصیات داده‌ها ارائه می‌دهند.

ماتریس همبستگی ویژگی‌ها

علاوه بر محاسبات آماری، ماتریس همبستگی ویژگی‌های مجموعه داده نیز مورد تجزیه و تحلیل قرار گرفته است که در شکل ۵ نمایش داده شده است. ماتریس همبستگی، ارتباط میان ویژگی‌های مختلف مجموعه داده را بررسی کرده و به شناسایی هرگونه ارتباط معنادار یا وابستگی بین متغیرها کمک می‌کند.

جدول. 2 محاسبه آماری ویژگی ها

شکل. 5 تجسم نقشه حرارتی همبستگی های ویژگی

مراجع

1. Jain D, Shah S, Mehta H et al (2021) A Machine Learning Approach to Analyze Marine Life Sustainability.

In: Proceedings of International Conference on Intelligent Computing, Information and Control

Systems. Springer, pp 619–632

2. Clark RM, Hakim S, Ostfeld A (2011) Handbook of water and wastewater systems protection. In: Protecting

Critical Infrastructure. Springer, pp 1–29. https:// doi. org/ 10. 1007/ 978-1- 4614- 0189-6

3. Hu Z, Zhang Y, Zhao Y et al (2019) A water quality prediction method based on the deep LSTM network

considering correlation in smart mariculture. Sensors 19:1420

4. Zhou J, Wang Y, Xiao F et al (2018) Water quality prediction method based on IGRA and LSTM.

Water 10:1148

5. Waqas M, Tu S, Halim Z et al (2022) The role of artificial intelligence and machine learning in wireless

networks security: principle, practice and challenges. Artif Intell Rev 55:5215–5261. https:// doi.

org/ 10. 1007/ s10462- 022- 10143-2

6. Halim Z, Waqar M, Tahir M (2020) A machine learning-based investigation utilizing the in-text features

for the identification of dominant emotion in an email. Knowl Based Syst 208:106443. https://

doi. org/ 10. 1016/j. knosys. 2020. 106443

7. Wu J, Wang Z (2022) A Hybrid Model for Water Quality Prediction Based on an Artificial Neural

Network, Wavelet Transform, and Long Short-Term Memory. Water 14:610

8. Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four Major South Korea’s Rivers

using deep learning models. Int J Environ Res Public Health 15:1322

9. Liu P, Wang J, Sangaiah AK et al (2019) Analysis and prediction of water quality using LSTM

deep neural networks in IoT environment. Sustainability 11:2058

10. Hmoud Al-Adhaileh M, Waselallah Alsaade F (2021) Modelling and prediction of water quality by

using artificial intelligence. Sustainability 13:4259

11. Bhardwaj D, Verma N (2017) Research paper on analysing impact of various parameters on water

quality index. Int J Adv Res Comput Sci 8(5):2496–498

12. Malek NHA, Wan Yaacob WF, Md Nasir SA, Shaadan N (2022) Prediction of Water Quality Classification

of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. Water 14:1067

13. Slatnia A, Ladjal M, Ouali MA, Imed M (2022) Improving prediction and classification of water

quality indices using hybrid machine learning algorithms with features selection analysis. In:

Online International Symposium on Applied Mathematics and Engineering (ISAME22), vol

1. ISAME22, Istanbul-Turkey, pp 16–17

14. Deng T, Chau K-W, Duan H-F (2021) Machine learning based marine water quality prediction for

coastal hydro-environment management. J Environ Manage 284:112051

15. Khullar S, Singh N (2022) Water quality assessment of a river using deep learning Bi-LSTM methodology:

forecasting and validation. Environ Sci Pollut Res 29:12875–12889

16. Abba SI, Pham QB, Saini G et al (2020) Implementation of data intelligence models coupled with ensemble

machine learning for prediction of water quality index. Environ Sci Pollut Res 27:41524–41539

17. Elbeltagi A, Pande CB, Kouadri S, Islam ARM (2022) Applications of various data-driven models

for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. Environ Sci

Pollut Res 29:17591–17605

18. Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM (2021) River water quality index prediction and

uncertainty analysis: A comparative study of machine learning models. J Environ Chem Eng 9:104599

19. Nosair AM, Shams MY, AbouElmagd LM et al (2022) Predictive model for progressive salinization

in a coastal aquifer using artificial intelligence and hydrogeochemical techniques: A case study

of the Nile Delta aquifer, Egypt. Environ Sci Pollut Res 29:9318–9340

20. Garabaghi FH, Benzer S, Benzer R (2021) Performance evaluation of machine learning models

with ensemble learning approach in classification of water quality indices based on different subset

of features. Res Square 1:1–35. https:// doi. org/ 10. 21203/ rs.3. rs- 876980/ v2

21. Hassan MM, Hassan MM, Akter L et al (2021) Efficient Prediction of Water Quality Index (WQI)

Using Machine Learning Algorithms. Hum Centric Intell Syst 1:86–97

22. Radhakrishnan N, Pillai AS (2020) Comparison of Water Quality Classification Models using

Machine Learning. In: 2020 5th International Conference on Communication and Electronics Systems

(ICCES). IEEE, pp 1183–1188

23. Khan MSI, Islam N, Uddin J et al (2021) Water quality prediction and classification based on principal

component regression and gradient boosting classifier approach. J King Saud Univ – Comput Inform

Sci 34(8):4773–4781. https:// doi. org/ 10. 1016/j. jksuci. 2021. 06. 003

24. Aldhyani THH, Al-Yaari M, Alkahtani H, Maashi M (2020) Water quality prediction using artificial

intelligence algorithms. Appl Bionics Biomech 2020:1–12. https:// doi. org/ 10. 1155/ 2020/ 66593 14

25. Khoi DN, Quan NT, Linh DQ et al (2022) Using Machine Learning Models for Predicting the

Water Quality Index in the La Buong River, Vietnam. Water 14:1552

26. Forests R, Breiman L (1999) Statistics Department University of California Berkeley. pp 1-29

27. Biau G (2012) Analysis of a random forests model. J Mach Learn Res 13:1063–1095

28. Wang S, Peng H, Liang S (2022) Prediction of estuarine water quality using interpretable machine

learning approach. J Hydrol 605:127320

29. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd

ACM sigkdd international conference on knowledge discovery and data mining. pp 785–794

30. Prakash R, Tharun VP, Devi SR (2018) A comparative study of various classification techniques

to determine water quality. In: 2018 Second International Conference on Inventive Communication

and Computational Technologies (ICICCT). IEEE, pp 1501–1506

31. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378

32. Zhou Y, Mazzuchi TA, Sarkani S (2020) M-adaboost-a based ensemble system for network intrusion

detection. Expert Syst Appl 162:113864

33. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is “nearest neighbor” meaningful? In:

International conference on database theory. Springer, pp 217–235

34. Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality

prediction. Chemosphere 249:126169

35. Halim Z, Rehan M (2020) On identification of driving-induced stress using electroencephalogram signals:

A framework based on wearable safety-critical scheme and machine learning. Inf Fusion 53:66–

79. https:// doi. org/ 10. 1016/j. inffus. 2019. 06. 006

36. Chen H, Huang JJ, McBean E (2020) Partitioning of daily evapotranspiration using a modified shuttleworth-

wallace model, random Forest and support vector regression, for a cabbage farmland. Agric

Water Manag 228:105923

37. Cheng Y, Peng J, Gu X et al (2020) An intelligent supplier evaluation model based on data-driven support

vector regression in global supply chain. Comput Ind Eng 139:105834

38. Liao Z, Li Y, Xiong W et al (2020) An In-Depth Assessment of Water Resource Responses to Regional

Development Policies Using Hydrological Variation Analysis and System Dynamics Modeling. Sustainability

12:5814

39. Tyagi S, Sharma B, Singh P, Dobhal R (2013) Water quality assessment in terms of water quality

index. Am J Water Resour 1:34–38

40. Shams MY, Tarek Z, Elshewey AM et al (2023) A Machine Learning-Based Model for Predicting

Temperature Under the Effects of Climate Change. In: Hassanien AE, Darwish A (eds) The Power

of Data: Driving Climate Change with Data Science and Artificial Intelligence Innovations. Springer

Nature Switzerland, Cham, pp 61–81

41. Elshewey AM, Shams MY, Elhady AM et al (2023) A Novel WD-SARIMAX Model for Temperature

Forecasting Using Daily Delhi Climate Dataset. Sustainability 15:757. https:// doi. org/ 10. 3390/ su150 10757

42. Tarek Z, Shams MY, Elshewey AM et al (2023) Wind Power Prediction Based on Machine Learning and

Deep Learning Models. Comput Mater Contin 74:715–732. https:// doi. org/ 10. 32604/ cmc. 2023. 032533

43. Elshewey AM, Shams MY, Tarek Z et al (2023) Weight Prediction Using the Hybrid Stacked-LSTM

Food Selection Model. Comput Syst Sci Eng 46:765–781. https:// doi. org/ 10. 32604/ csse. 2023. 034324

# **نویسندگان و وابستگی‌های سازمانی**

محمود ی. شمس¹، احمد م. الشیوی²، السید م. الکناوی³، عبدالحمید ابراهیم⁴، فاطمه م. طلعت¹,⁵، زهراء طارق⁶

Mahmoud Y. Shams

mahmoud.yasin@ai.kfs.edu.eg

Ahmed M. Elshewey

ahmed.elshewey@fci.suezuni.edu.eg

El‑Sayed M. El‑kenawy

skenawy@ieee.org

Abdelhameed Ibrahim

afai79@mans.edu.eg

Fatma M. Talaat

fatma.nada@ai.kfs.edu.eg

Zahraa Tarek

zahraatarek@mans.edu.eg

1. دانشکده هوش مصنوعی، دانشگاه کفرالشیخ، کفرالشیخ 33516، مصر
2. دانشکده رایانه و اطلاعات، گروه علوم رایانه، دانشگاه سوئز، سوئز، مصر
3. گروه ارتباطات و الکترونیک، مؤسسه عالی مهندسی و فناوری دلتا، منصوره 35111، مصر
4. گروه مهندسی کامپیوتر و سیستم‌های کنترل، دانشکده مهندسی، دانشگاه منصوره، منصوره 35516، مصر
5. دانشکده علوم و مهندسی کامپیوتر، دانشگاه جدید منصوره، منصوره 35712، مصر
6. دانشکده رایانه و اطلاعات، گروه علوم رایانه، دانشگاه منصوره، منصوره 35561، مصر