نشریه آبیاری و زهکشی ایران
شماره۲، جلد ۱۸، خرداد - تیر ۱۴۰۳، ص. ۳۱۷-۳۰۷

Iranian Journal of Irrigation and Drainage
No.2, Vol. 18, Jun.-Jul. 2024, p. 307-317

مقاله علمی-پژوهشی

پیش‌بینی غلظت نیترات در آب‌های زیرزمینی محدوده شرق استان مازندران با استفاده از الگوریتم‌های محاسباتی نرم

فرحناز دوستعلی‌زاده[۱]، مجتبی خوش‌روش[۲*]، رامین فضل‌اولی[۳]، محمدمهدی باطنی[۴]

تاریخ دریافت: ۱۴۰۲/۰۹/۱۰ تاریخ پذیرش: ۱۴۰۳/۰۳/۱۷

چکیده

با توجه به اهمیت آب شیرین برای حیات بشر و آسیب‌پذیری منابع آب زیرزمینی در برابر انواع آلودگی‌ها و امکان انتقال آلاینده‌ها به سایر منابع آب-های سطحی و زیرزمینی و همچنین، قرار داشتن کشور ایران در کمربند خشک و نیمه‌خشک، محافظت از این عنصر با ارزش و کمیاب بسیار ضروری بوده و پایش مداوم و مستمر آن باید از اولویت‌های مدیریت منابع آب قرار گیرد. از این رو، در پژوهش حاضر به آلودگی نیترات در دشت‌های شرق استان مازندران و بررسی مسائل مربوطه پرداخته شد و مدلی کارآمد و بهینه جهت پیش‌بینی غلظت نیترات ارائه شد. در انجام این پژوهش به مقایسه‌ی سه مدل یادگیری ماشین درخت تصمیم، رگرسیون لجستک و شبکه عصبی مصنوعی پرداخته شد. داده‌های فیزیکی و شیمیایی اندازه‌گیری شده طی سال‌های ۱۳۶۴ تا ۱۳۹۹ استفاده شده و به عنوان متغیرهای ورودی مدل‌ها قرار داده شد. متغیرها شامل دما، سطح ایستابی، pH، EC، HCO_3^-، CL، SO_2^{4-}، Na^+، K^+، Mg^{2+}، Ca^{2+}، TH و TDS بوده و میزان آلودگی نیترات آب زیرزمینی با تقسیم ۷۰ درصد داده‌ها به عنوان آموزش و ۳۰ درصد به عنوان آزمون، پیش‌بینی شده و از شاخص‌های R^2، RMSE، NSE و PBIAS برای ارزیابی مدل‌ها استفاده شد. نتایج به دست آمده حاکی از آن بود که مدل درخت تصمیم با اختلاف زیاد نسبت به دو مدل دیگر بهترین عملکرد را داشته (R^2 = ۰/۹۵۷ و RMSE=۰/۲۹۷، NSE =۰/۹۵ و Testing acc =۰/۹۰۷) و پس از آن رگرسیون لجستیک و شبکه عصبی مصنوعی با عملکردهای به مراتب ضعیف‌تری نسبت به مدل درخت تصمیم قرار داشتند. پیشنهاد می‌شود آزمایش با مدل‌های دیگر یادگیری ماشین و تغییر قراردادن متغیرهای ورودی و اضافه کردن چند متغیر دیگر از جمله کاربری اراضی و بارندگی انجام و نتایج با پژوهش حاضر مقایسه گردند.

واژه‌های کلیدی: داده‌کاوی، درخت‌های تصمیم، رگرسیون لجستیک، شبکه عصبی مصنوعی

مقدمه

آب زیرزمینی یکی از منابع مهم و با ارزش برای بهره‌برداری و استفاده از لحاظ کشاورزی، شرب، صنعت و غیره، بالاخص در مناطق خشک و نیمه خشک می‌باشد که درصد بالایی از آب کشاورزی و شرب کشور ایران را تامین می‌کند (IMOF, 2014؛ حسینی وردنجانی و همکاران، ۱۴۰۲). نشاط و همکاران گزارش کردند که با توجه به مصرف بیش از اندازه و آلاینده‌های وارد شده به آب زیرزمینی و فقدان منبع مناسب برای جایگزینی آن، احتمالا یکی از مشکلات اساسی در دراز مدت، مواجهه با کمبود آن خواهد بود (Neshat et al., 2014). به دلیل شرایط آب و هوایی کشور ایران و روند نزولی متوسط بارش سالانه، اهمیت این منبع طبیعی و محدود دو چندان بوده و محافظت آن از انواع آلودگی‌های محلول و نامحلول در آب بسیار ضروری است. نیترات از جمله موادی است که به آسانی قابلیت حل شدن با آب را داشته و مخصوصا در زمین‌های کشاورزی حاوی کودهای نیتروژنه، از خلل و فرج خاک عبور کرده و به آب زیرزمینی راه پیدا می‌کند. از جمله راه‌های افزایش میزان نیترات در آب زیرزمینی می‌توان به دفع فاضلاب و بهداشت در محل اشاره کرد که با مشخص شدن منشا آن کمک شایانی به مدیریت این آلودگی خواهد شد. از آن جا که این ماده قابلیت ماندگاری بالایی در خاک

۱- دانش آموخته کارشناسی ارشد، گروه مهندسی آب، ، دانشکده مهندسی زراعی، دانشگاه علوم کشاورزی و منابع طبیعی ساری، ساری، ایران
۲- دانشیار، گروه مهندسی آب، دانشکده مهندسی زراعی، دانشگاه علوم کشاورزی و منابع طبیعی ساری، ساری، ایران.
۳- دانشیار، گروه مهندسی آب، دانشکده مهندسی زراعی، دانشگاه علوم کشاورزی و منابع طبیعی ساری، ساری، ایران.
۴- پژوهشگر، انستیتوی مطالعات پیشرفته (IUSS)، پاویا، ایتالیا
(*- نویسنده مسئول: Email: m.khoshravesh@sanru.ac.ir)

داشته، شناخت محل تجمع آن و چگونگی افزایش و انتقال آن به دیگر آبخوان‌ها ضروری به نظر می‌رسد. به طور خاص، در دشت‌های ممنوعه، این موضوع مهم‌تر جلوه می‌کند. زیرا افت شدید سطح آب، می‌تواند این دشت‌ها را از حالت بحرانی به فوق بحرانی تبدیل کرده، بنابراین اجازه تاسیس چاه جدید و بهره‌برداری بیش از حد داده نشده و این مناطق تحت حفاظت سازمان‌های مربوطه می‌باشند.

نیترات (Nitrate) ترکیبی است که توسط میکروارگانیزم‌های آب و خاک در اثر ترکیب نیتروژن با اکسیژن یا اوزن به‌طور طبیعی ایجاد می‌گردد. همچنین یک یون چند اتمی با فرمول شیمیایی NO_3^- است که به یون اسید نیتریک یا یون نیترات نیز معروف می‌باشد. این ترکیب در گروه ترکیبات غیر آلی بوده که به عنوان نیترات‌های غیر فلزی شناخته می‌شوند و به عنوان بزرگ‌ترین گروه آنیون حاوی اکسیژن هستند. تقریبا همه‌ی نیترات‌ها در آب محلول می‌باشند. نیترات‌ها به خودی خود سمی نیستند اما از آن‌جا که محصولات متابولیکی آن‌ها از طریق نیتریت به ترکیبات N-nitro (این ترکیبات قابلیت انفجاری بالایی دارند) تبدیل می‌شوند، در علم سم‌شناسی بسیار مورد بحث قرار می‌گیرند (Gangolli et al., 1994; WHO, 1995). تصور می‌شود که نیترات به دلیل تولید بالقوه نیتروزامین‌های سرطان‌زا تحت شرایط خاصی مانند معده‌ی اسیدی مضر است. نیتروزامین‌ها با سرطان مری، سرطان معده، سرطان روده بزرگ و سایر تومورها مرتبط هستند (Ma et al., 2018). نیترات از جمله موادی است که به آسانی قابلیت حل شدن با آب را داشته و مخصوصا در زمین‌های کشاورزی حاوی کودهای نیتروژنه، از خلل و فرج خاک عبور کرده و به آب زیرزمینی راه پیدا می‌کند. از جمله راه‌های افزایش میزان نیترات در آب زیرزمینی می‌توان به دفع فاضلاب و بهداشت در محل اشاره کرد که با مشخص شدن منشا آن کمک شایانی به مدیریت این آلودگی خواهد شد. از آن جا که این ماده قابلیت ماندگاری بالایی در خاک داشته، شناخت محل تجمع آن و چگونگی افزایش و انتقال آن به دیگر آبخوان‌ها ضروری به نظر می‌رسد.

حسینی و مهجوری با ترکیب چند روش با SVR (FNN-SVR, MLP-SVR, SVR-GA) به ارزیابی این روش در مکان‌یابی غلظت نیترات در آکیفر شهر کرج استان البرز پرداختند. آن‌ها از داده‌های غلظت نیترات به‌عنوان فاکتور ورودی استفاده کردند. نتایج نشان داد که مدل FNN-SVR بیشترین هماهنگی را با داده‌های اندازه‌گیری شده‌ی مزرعه‌ای داشت (Hosseini and Mahjouri, 2014) (SE)= 0.31). رخشاد و همکاران به ارزیابی خطر آلودگی نیترات آب‌های زیرزمینی در مشهد پرداختند. چهار مدل یادگیری ماشین برای ارزیابی احتمال آلودگی آب‌های زیرزمینی از جمله مدل GLM، BRT و SVM استفاده و اعتبار هر مدل با منحنی مشخصه AUC ارزیابی شد. برای ترکیب مدل‌ها، حداقل آستانه ۸۰ درصد در نظر گرفته شد. نتایج نشان داد که دقت چهار مدل بین ۸۰ تا ۹۰ درصد متغیر است؛ بنابراین تمامی مدل ها در بخش ترکیب در نظر گرفته شدند. با توجه به نتایج، روش گروهی می‌تواند کارایی خوبی را ارائه دهد (Rokhshad et al., 2021).

آویس و همکاران در پاکستان از سه الگوریتم یادگیری ماشین SVM، Multivariate Discriminant Analysis (MDA) و BRT برای ارزیابی ریسک آلودگی نیترات استفاده کردند. مدل‌ها از طریق آزمایش های واسنجی، با استفاده از ناحیه تحت روش منحنی مشخصه عملکرد گیرنده (AUC)، که در آن حداقل مقدار آستانه AUC برابر ۸۰ درصد به‌دست آمد، واسنجی و اعتبارسنجی شدند. نتایج نشان داد که دقت مدل‌ها در محدوده ۰/۸۷ تا ۰/۸۲ است. نقشه نهایی خطر آلودگی آب های زیرزمینی نشان داد که ۳۴ درصد از منطقه نسبت به آلودگی آب‌های زیرزمینی نسبتا آسیب‌پذیر بوده و ۱۳ درصد از منطقه در معرض خطر آلودگی بالای آب های زیرزمینی است (Awais et al., 2021). الزاین و همکاران با استفاده از مدل‌های پیشرفته چندگانه یادگیری ماشین (ML) شبکه‌های عصبی پایه شعاعی (RBNN)، SVR و RFR به تعیین دقیق ترین عملکرد برای ارزیابی آسیب‌پذیری آلودگی آب‌های زیرزمینی پرداختند. هشت عامل آسیب پذیری DRASTIC-L بر اساس مدل اصلاح شده‌ی DRASTIC (MDM) رتبه‌بندی شده و به عنوان داده‌های ورودی استفاده شدند. شاخص آسیب پذیری تنظیم شده (AVI) با مقادیر نیترات به عنوان داده های خروجی برای فرایند مدل‌سازی استفاده شد. عملکرد سه مدل با استفاده از معیارهای عملکرد آماری مقادیر MAE، RMSE، R^2 و ROC/AUC تایید شد. مدل RFR گروهی بالاترین عملکرد را در مقایسه با مدل های مستقل SVR و RBNN نشان داد. به‌طور خاص، مجموعه RFR به‌دلیل انعطاف‌پذیری و استحکام، تمام راه‌حل‌های امیدوارکننده را در طول عملکرد مدل حفظ کرد و نقشه آسیب‌پذیری به‌دست‌آمده توسط مدل RFR برای پیش‌بینی آسیب‌پذیرترین مناطق در برابر آلودگی دقیق‌تر بود. نتیجه‌گیری شد که مجموعه RFR ابزاری قوی برای افزایش ارزیابی آسیب‌پذیری آلودگی آب‌های زیرزمینی بوده و می‌تواند به ایمنی محیط زیست در برابر آلودگی آب‌های زیرزمینی کمک کند (Elzain et al., 2022). باند و همکاران در حوضه‌ی آبخیز مرودشت، به مدل‌سازی غلظت نیترات آب‌های زیرزمینی فضایی توسط الگوریتم‌های SVR, Random Forest, cubist و Baysia-ANN پرداختند. آن‌ها از ۱۱ متغیر مستقل موثر بر نیترات از جمله ارتفاع، شیب، انحنای پلان، انحنای نیم رخ، بارندگی، عمق پیزومتریک، فاصله از رودخانه، فاصله از منطقه‌ی مسکونی، سدیم (Na)، پتاسیم (K) و شاخص رطوبت توپوگرافی (TWI) در منطقه‌ی مورد مطالعه استفاده کرده و نتایج نشان داد که الگوریتم RF با مقادیر (R^2 =

0.89, RMSE = 4.24, NSE = 0.87) از سه الگوریتم دیگر عملکرد بهتری داشته است (Band et al., 2020). ساجدی حسینی و همکاران در دشت لنجانات اصفهان، برای ارزیابی آسیب‌پذیری آب‌های زیرزمینی به آلودگی نیترات از روش‌های درخت‌های تصمیم تقویت شده، تجزیه و تحلیل تفکیک چند متغیره، ماشین‌بردار پشتیبان استفاده کرده و جهت صحت‌سنجی مدل نیز از روش منحنی مشخصه عملکرد گیرنده بهره جستند. طبق نتایج، تمامی روش‌ها با AUC= 0.81 to 0.85 و MSE=0.16 to 0.23 از عملکرد مناسبی برخوردار بودند (Sajedi-Hosseini et al., 2018). سونگ هه و همکاران غلظت نیترات آب زیرزمینی کم عمق در منطقه Yinchuan در دشت Yinchuan مرکزی در طول سال های ۲۰۰۰، ۲۰۰۵، ۲۰۱۰ و ۲۰۱۵ را با استفاده از جنگل تصادفی مدل‌سازی کردند. عوامل محیطی چندگانه به عنوان متغیرهای پیش بینی در نظر گرفته شدند. اهمیت نسبی این عوامل نیز با استفاده از مدل ساخته شده محاسبه شد. از روش های سنجش از دور و GIS برای گردآوری عوامل محیطی مختلف برای تولید مجموعه های آموزشی و آزمایشی برای آموزش و اعتبارسنجی مدل جنگل تصادفی استفاده شد. شاخص‌های میانگین خطای مطلق (MAE)، ریشه میانگین مربعات خطا (RMSE) و ضریب تعیین (R^2) بین غلظت نیترات آب زیرزمینی مشاهده شده و پیش‌بینی‌شده برای اندازه‌گیری عملکرد مدل استفاده شد. طبق نتایج، مدل جنگل تصادفی برای پیش بینی نیترات آب زیرزمینی به خوبی انجام شد. اهمیت نسبی متغیرهای پیش‌بینی‌کننده محاسبه‌شده توسط مدل نشان داد که نیترات آب‌های زیرزمینی عمدتا تحت تاثیر فاصله تا رودخانه زرد، عناصر هواشناسی (بارش، تبخیر و میانگین دمای هوا) و ارتفاع سطح آب قرار می گیرد. علاوه بر این، زمین شهری و زراعی دو نوع کاربری/پوشش زمین بودند که عمدتا بر غلظت نیترات آب زیرزمینی در منطقه یین چوان تاثیر گذاشتند، که در نتیجه گسترش شدید زمین شهری از سال ۲۰۰۰ تا ۲۰۱۵، زمین شهری از زمین های زراعی تاثیرگذارتر بود (He et al., 2022). گارسیا دل تورو و همکاران در مورسیای اسپانیا، به بررسی کیفیت آب زیرزمینی این منطقه پرداختند. برای کنترل و پیش‌بینی کیفیت آب زیرزمینی این منطقه، دو مدل یادگیری ماشین (Naïve-Bayes و Decision-tree) پیشنهاد شد. این مدل‌ها به قدرت محاسباتی زیادی نیاز نداشتند و با استفاده از ابزار KNIME (KoNstanz Information MinEr) از تعداد کمتری داده‌ها توسعه یافتند. دقت آن‌ها توسط confusion matrix مربوطه آزمایش شد و دقت بالایی را در هر دو مدل ارائه داد. نتایج به‌دست‌آمده نشان داد که کیفیت آب زیرزمینی در پهنه‌های شمال و غرب بالاتر است (García-del-Toro et al., 2022). غلامی و بوجی از سه روش یادگیری ماشینی شامل شبکه عصبی عمیق (DNN)، تقویت گرادیان شدید (EGB) و رگرسیون خطی چندگانه (MLR) برای پیش بینی آلودگی نیترات در آب های زیرزمینی شمال ایران (دشت مازندران) استفاده کردند. میانگین غلظت نیترات در ۲۵۰ حلقه چاه پیزومتریک به عنوان متغیر خروجی در نظر گرفته شده و عوامل موثر بر کیفیت آب زیرزمینی (عمق آب زیرزمینی، قابلیت انتقال سفره های زیرزمینی، بارش، تبخیر، فاصله از منابع آبی و دریای خزر، فاصله از صنایع و مراکز مسکونی، تراکم جمعیت، توپوگرافی و بهره‌برداری از آب های زیرزمینی) به‌عنوان متغیرهای ورودی در یک سفره آبرفتی در نظر گرفته شدند. نتایج مراحل آموزش و آزمایش نشان داد که روش EGB بالاترین عملکرد را در پیش بینی غلظت نیترات به‌دلیل کمترین مقادیر خطا و بیشترین همبستگی بین مقادیر اندازه‌گیری شده و پیش‌بینی‌شده غلظت نیترات داشت (آموزش R-sqr = 0.98، Nash-Sutcliffe (NSE) = 0.98 و آزمون R-sqr = 0.86، NSE = 0.84). همچنین نتایج حاکی از آن بود که فاکتورهای فاصله از صنایع، تراکم جمعیت، عمق آب زیرزمینی و میزان تبخیر از مهم ترین عوامل مؤثر بر غلظت نیترات در آب های زیرزمینی بود. در نهایت، مدل EGB آزمایش شده و ابزار سیستم اطلاعات جغرافیایی (GIS) برای تهیه نقشه آلودگی آب های زیرزمینی نیترات در منطقه مورد مطالعه استفاده شد. ارزیابی عملکرد نقشه حاصل با مقایسه مقادیر پیش‌بینی‌شده و اندازه‌گیری شده، دقت خوبی را نشان داد (R-sqr = 0.8)(Gholami& Booij, 2022). اجلیل و همکاران پنج مدل جدید یادگیری ماشین ترکیبی/گروهی (ML) به نام‌های DRASTIC-Random Forest (RF)، DRASTIC Support Vector Machine (SVM)، DRASTIC-Multilayer Perceptron (MLP)، DRASTIC-RF-SVM، و DRASTIC-RF-MLP را برای ارزیابی آلودگی آب‌های زیرزمینی در حوضه Saiss، در مراکش توسعه دادند. عملکرد این مدل‌ها با استفاده از منحنی مشخصه عملیاتی گیرنده (منحنی ROC)، precision وaccuracy ارزیابی شد. بر اساس نتایج روش منحنی ROC، نشان داده شد که استفاده از مدل‌های یادگیری ماشین ترکیبی/ گروهی (ML) عملکرد الگوریتم‌های یادگیری ماشین فردی (ML) را بهبود بخشید. در واقع، مقدار AUC، DRASTIC اصلی ۰/۵۱ بود. علاوه بر این، هر دو مدل هیبریدی/گروهی، DRASTIC-RF-MLP (AUC= 0.953) و DRASTIC-RFSVM(AUC = 0.901)، بهترین دقت را در بین مدل‌های دیگر به‌دست آورده و پس از آن DRASTIC-RF(AUC 0.852)، DRASTIC-SVM (AUC = 0.802) و DRASTIC-MLP (AUC = 0.763) قرار داشت (Ijlil et al., 2022).

با توجه به مسائل مطرح شده، به نظر می‌رسد مطالعه‌ی منابع آبی موجود و حفظ کیفیت آن‌ها، حائز اهمیت می‌باشد. یکی از راه‌های پژوهش و بررسی این منابع، مدل‌سازی بر اساس داده‌های موجود و در نتیجه تفسیر نتایج حاصل از مدل می‌باشد. هنگامی که بتوان به

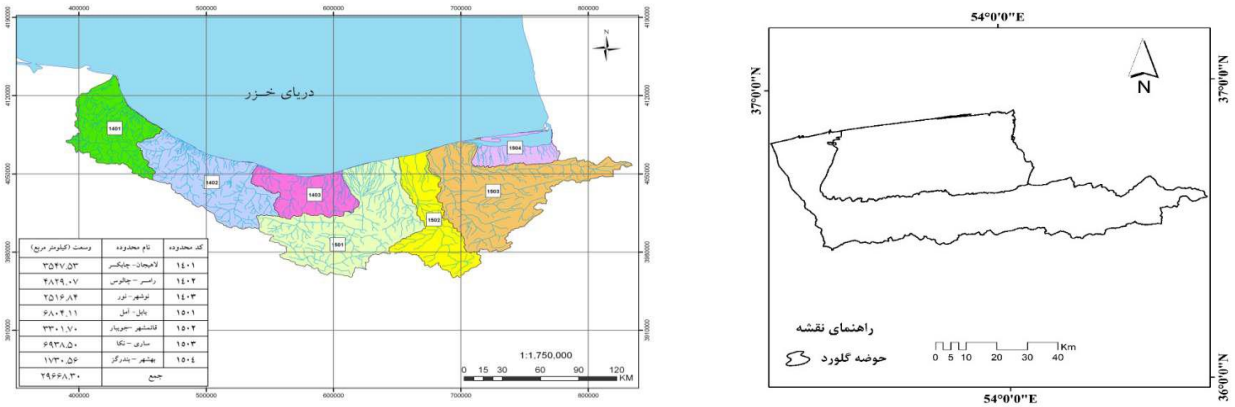
یک مدل بهینه دست یافت، می‌توان از نتایج آن نیز در مسائل مدیریتی و افزایش بهره‌وری منابع آبی استفاده نمود. با توجه به رشد و پیشرفت روزافزون در علم کامپیوتر و زبان‌های برنامه‌نویسی، انجام پژوهش‌های مختلف در حوزه‌ی محیط زیست و منابع آب، با افزایش دقت و سهولت همراه بوده و امکان بررسی سناریوهای مختلف در هر حیطه را به محققین می‌دهد. برای انجام این پژوهش از روش‌های نوین یادگیری ماشین (به کمک زبان برنامه نویسی پایتون) استفاده و میزان دقت هر یک از زیرشاخه‌ها نیز اندازه‌گیری شد.

مواد و روش‌ها

این پژوهش در محدوده‌ی مطالعاتی ساری-نکا و بهشهر-بندرگز استان مازندران انجام شد (شکل ۱). محدوده‌ی مطالعاتی ساری-نکا بین طول‌های جغرافیایی ´۳۴ °۵۴ تا ´۴۴ °۵۴ شرقی و ´۵۶ °۳۵ تا ´۵۲ °۳۶ شمالی در بخش خاوری حوضه‌های درجه دو بین رودخانه هراز و سفید رود و رودخانه‌ی هراز و رودخانه‌های بین هراز و قره‌سو و در جنوب محدوده‌ی بهشهر-بندرگز قرار داشته و مساحت آن نزدیک به ۶۹۳۸/۵ کیلومتر مربع می‌باشد که ۹۷۲/۸۷ کیلومتر مربع آن دشت و بقیه (۵۹۶۵/۶ کیلومتر مربع) شامل ارتفاعات می‌باشد. بالاترین نقطه‌ی ارتفاعی این محدوده ۳۸۳۶ متر و پایین‌ترین نقطه‌ی ارتفاعی آن با ۲۶- متر از سطح دریای آزاد در خروجی حوضه قرار دارد. از مراکز جمعیتی مهم این محدوده‌ی مطالعاتی می‌توان شهرهای ساری و نکا را نام برد.

در این محدوده، تعداد ۱۸۶۹۵ حلقه چاه با برداشت سالانه ۱۹۱/۰۳ میلیون متر مکعب، ۳۰۷۵ دهنه چشمه با تخلیه‌ی سالانه‌ی ۹۸/۲۸ میلیون متر مکعب و ۱۵ رشته قنات با تخلیه‌ی ۴/۹۵ میلیون متر مکعب در سال شناسایی شده است. حجم کل تخلیه منابع آب زیرزمینی در این محدوده‌ی مطالعاتی برابر با ۲۹۴/۲۶ میلیون متر مکعب در سال است که ۱۹۰/۹۷ میلیون متر مکعب آن مربوط به دشت ساری-نکا و مابقی مربوط به ناحیه ارتفاعات مشرف به آن می‌باشد. در سطح آبخوان آبرفتی دشت ساری-نکا تعداد ۱۶۹۴۷ حلقه چاه، ۱۴ دهنه چشمه و ۱۵ رشته قنات با مجموع برداشت و تخلیه‌ی سالانه‌ی ۱۸۵/۸۳ میلیون متر مکعب شناسایی شده است.

حجم کل آب مصرفی در سطح محدوده مطالعاتی ساری-نکا برابر با ۷۱۶/۷۵ میلیون متر مکعب در سال است که ۶۵۳/۳۱ میلیون متر مکعب آن در بخش کشاورزی و ۵۵/۷۲ میلیون متر مکعب آن در بخش شرب و مابقی در بخش صنعت استفاده می‌گردد.



شکل ۱- موقعیت محدوده‌های مطالعاتی

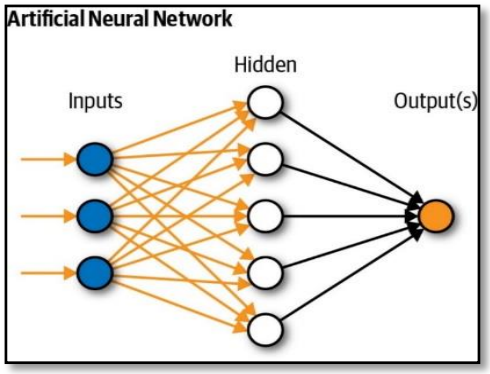
محدوده‌ی مطالعاتی بهشهر-بندرگز نیز با کد ۱۵۰۴ بین طول‌های جغرافیایی ´۱۹ °۵۳ تا ´۵۰ °۵۴ شرقی و ´۳۷ °۳۶ تا ´۵۶ °۳۶ شمالی ۱۵۰۴ و در شرقی‌ترین بخش حوضه‌ی تلفیق مازندران و شرق گیلان قرار دارد. وسعت محدوده‌ی مطالعاتی بهشهر-بندرگز ۱۷۳۰/۵ کیلومترمربع می‌باشد که ۷۵۲/۲۵ کیلومتر مربع آن مربوط به دشت بهشهر-بندرگز و ۴۴۰/۷۵ کیلومتر مربع مربوط به ارتفاعات و ۵۳۷/۵۶ کیلومتر مربع مربوط به خلیج گرگان می‌باشد. بالاترین نقطه‌ی ارتفاعی محدوده ۲۳۲۱ و پایین‌ترین نقطه‌ی آن با ارتفاع ۲۶- متر از سطح دریای آزاد در خروجی حوضه قرار دارد. از مراکز جمعیتی مهم این محدوده می‌توان شهرهای بهشهر و بندرگز را نام برد.

در انجام این پژوهش از الگوریتم‌های متفاوتی بر اساس روش‌های محاسباتی نرم استفاده شد که شامل رگرسیون لجستیک (LR) و درخت‌های تصمیم (DT) و شبکه‌ی عصبی مصنوعی (ANN)می‌باشند. در این مطالعه عملکرد روش‌های محاسباتی نرم مختلف نسبت به یکدیگر سنجیده شد. داده‌های مورد استفاده شامل خصوصیات فیزیکی و شیمیایی (آنیون‌ها و کاتیون‌ها) مختلف بوده که طی ۳۰ سال اندازه‌گیری شده و به‌عنوان داده‌های ورودی استفاده شد. این داده‌ها شامل دما، سطح ایستابی، pH، EC، HCO_3^-، CL^-، SO_2^{4-}

Na^+، K^+، Mg^{2+}، Ca^{2+}، TH، TDS و NO_3 بوده و میزان آلودگی نیترات آب زیرزمینی پیش‌بینی شد. در طول بررسی روش‌های مختلف یاد شده، از تکنیک Cross-Validation جهت انتخاب تصادفی از همه‌ی قسمت‌های داده‌های ورودی استفاده شده و ۷۰ درصد از داده‌ها به عنوان Train (آموزش) و ۳۰ درصد به عنوان Test (آزمون) استفاده شد.

شبکه عصبی مصنوعی(Artificial Neural Networks – ANN)

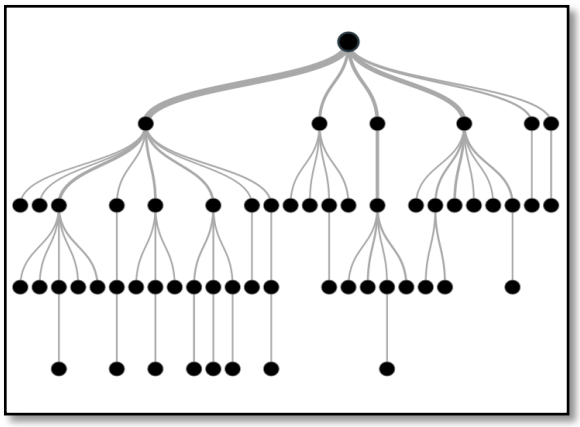
ANN یک سیستم مدل‌سازی الهام گرفته از شبکه عصبی انسان است که امکان یادگیری از یک پدیده فیزیکی یا توصیف یک فرایند تصمیم‌گیری توسط یک مثال را دارا بوده و قادر است روابط تجربی بین متغیرهای مستقل و وابسته برقرار کند و اطلاعات و دانش را از مجموعه داده‌های نماینده استخراج کند. شبکه‌های عصبی مصنوعی از لایه‌ای از گره‌های ورودی و لایه‌ای از گره‌های خروجی تشکیل شده‌اند که توسط یک یا چند لایه گره پنهان به هم متصل شده‌اند. لایه ورودی گره‌ها با توابع فعال‌سازی، اطلاعات را به گره‌های لایه پنهان منتقل کرده و گره‌های لایه پنهان بسته به شواهد ارائه شده فعال شده یا غیرفعال می‌مانند. لایه‌های پنهان توابع وزن‌دهی را به شواهد اعمال می‌کنند و زمانی که مقدار یک گره خاص یا مجموعه‌ای از گره‌ها در لایه‌ی پنهان به یک آستانه‌ی مشخص رسید، یک مقدار به یک یا چند گره در لایه خروجی منتقل می‌شود (شکل ۲).



شکل ۲- نمایی شماتیک از شبکه عصبی مصنوعی

درخت تصمیم (Decision Tree)

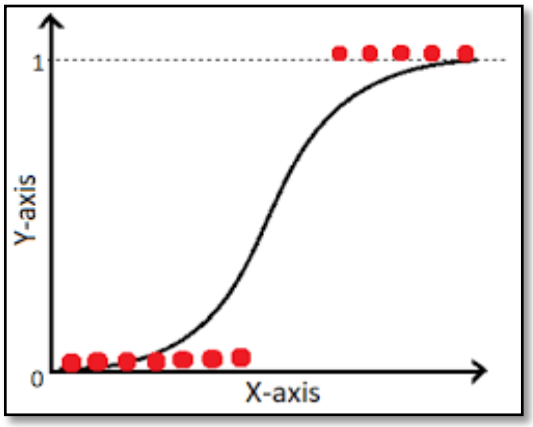
درخت تصمیم نوعی یادگیری ماشین نظارت شده است که برای طبقه‌بندی یا پیش‌بینی بر اساس نحوه‌ی پاسخ به مجموعه سوالات قبلی استفاده می‌شود. به این معنی که بر روی مجموعه‌ای از داده‌ها که شامل طبقه بندی مورد نظر است، آموزش دیده و آزمایش می شود. هر چه درخت عمیق‌تر باشد، قوانین تصمیم‌گیری پیچیده‌تر و مدل مناسب‌تر است (شکل ۳).



شکل ۳- نمایی شماتیک از درخت تصمیم

رگرسیون لجستیک (Logistic Regression)

رگرسیون لجستیک (Logistic Regression) یکی از الگوریتم‌های یادگیری ماشین است. این الگوریتم برای مسائل طبقه‌بندی (Classification) استفاده می‌شود که در آن متغیر وابسته‌ی گسسته مطرح می‌شود. به عبارت دیگر، رگرسیون لجستیک یک روش تحلیل آماری برای پیش بینی یک نتیجه باینری، مانند بله یا خیر، بر اساس مشاهدات قبلی یک مجموعه داده است. یک مدل رگرسیون لجستیک با تجزیه و تحلیل رابطه بین یک یا چند متغیر مستقل موجود، یک متغیر داده وابسته را پیش‌بینی می‌کند. در رگرسیون لجستیک، مستقیما مانند رگرسیون خطی، یک خط مستقیم با داده‌های خود مطابقت داده نمی‌شود. در عوض، یک منحنی S شکل به نام Sigmoid برای مشاهدات خود قرار داده می‌شود. این الگوریتم به راحتی قابل درک و قابل تفسیر بوده و می‌تواند نتایج بسیار خوبی ارائه دهند (شکل ۴).



شکل ٤- منحنی رگرسیون لجستیک

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i \quad (۱)$$

معادله (۱) : معادله رگرسیون لجستیک

Y_i: متغیر وابسته

β_0: عرض از مبدا

β_1: ضریب شیب

X_1: متغیر مستقل

ε_i: عبارت خطای تصادفی

فرایند مدل‌سازی

در ابتدا، برای قابل استفاده کردنِ داده‌ها برای فرایند کدنویسی، با تکنیک‌های مربوطه، داده‌ها پیش‌پردازش و آماده شدند. این تکنیک‌ها شامل پاک کردن داده‌های فاقد ارزش و پرت، رسم نحوه توزیع آن‌ها و در نهایت بالانس داده‌ها با تکنیک Resampling بودند. در مرحله‌ی بعد، برای ساخت مدل و تقسیم داده‌ها به دو قسمت آموزش و آزمون و ارزیابی مدل روی داده‌های کاملا جداگانه، از تکنیک Validation Hold Out استفاده شد. سپس، مرحله‌ی Feature Selection، برای مشخص کردن اهمیت و نقش هر ویژگی (ورودی) در میزان نیترات انجام شد. پس از آن، برای مدل‌سازی از ویژگی‌های موثرِ به دست آمده در مرحله‌ی قبل استفاده شده و به عنوان پارامترهای ورودی نهایی به مدل داده شدند. داده‌ها به دو قسمت آموزش و آزمون تقسیم شده و فرایند مدلسازی انجام شد. در انتها نیز،پس از ارزیابی مدل به دست آمده، تکنیک‌های مربوط به بهینه‌سازی پارامترهای پراهمیت در ساخت مدل انجام شده و مدل با مقادیر بهینه پارامترها بازسازی شد.

شاخص‌های مورد استفاده

شاخص‌های مورد استفاده برای صحت‌سنجی مدل‌ها به شرح زیر می‌باشند (تین بوی و همکاران، ۲۰۲۰):

معادله‌ی (۲)، ضریب تعیین (R^2-Score):

$$R^2 = \left(\frac{\sum_{i=1}^{n}(K_{Predicted} - \bar{K}_{Predicted})(K_{Measured} - \bar{K}_{Measured})}{\sqrt{\sum_{i=0}^{n}(K_{Predicted} - \bar{K}_{Predicted}) * \sum_{i=1}^{n}(K_{Measured} - \bar{K}_{Measured})}}\right) \quad (۲)$$

که در آن:

R^2= ضریب تعیین

$K_{predicted}$ = مقادیر پیش‌بینی شده

$\bar{K}_{predicted}$ = میانگین مقادیر پیش‌بینی شده

$K_{measured}$ = مقادیر اندازه‌گیری شده

$\bar{K}_{measured}$ = میانگین مقادیر اندازه‌گیری شده

معادله‌ی (۳)، میانگین مربع خطاها (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(K_{Predicted} - \bar{K}_{Predicted})^2} \quad (۳)$$

معادله‌ی (۴)، ضریب کارایی نش–ساتکلیف(NSE):

$$NSE = 1 - \frac{\sum_{i=1}^{n}(K_{Predicted} - K_{Measured})^2}{\sum_{i=1}^{n}(K_{Predicted} - \bar{K}_{Predicted})^2} \quad (۴)$$

معادله‌ی (۵)،درصد بایاس (Pbias):

$$PBIAS = \left(\frac{\sum_{i=1}^{n}(K_{Measured} - K_{Predicted})}{\sum_{i=1}^{n} K_{Predicted}}\right) * 100 \quad (۵)$$

نتایج و بحث

هدف از این پژوهش ایجاد یک مدل پیش‌بینی قابل اعتماد برای پیش‌بینی آلودگی نیترات در شرق استان مازندران است. برای نیل به این هدف، اعتبارسنجی مدل‌های پیشنهادی ANN، LR و DT با انجام آزمون همبستگی بر روی نتایج حاصل از آن‌ها و داده‌های واقعی موجود در طول سال‌های ۱۳۶۴ تا ۱۳۹۹ اجرا شده و نتایج آن ارائه شده است. برای این اعتبارسنجی از آزمون‌های R^2، RMSE، NSE و PBIAS استفاده شد. در این مطالعه از سیزده پارامتر موجود، برای ساخت مدل‌ها استفاده شد و نیترات به عنوان متغیر هدف (به عنوان دو حالت آلودگی یا غیرآلودگی) در نظر گرفته شد. پیش از شروع فرایند کدنویسی برای مدل‌سازی، داده‌ها توسط مرحله‌ی پیش پردازش از لحاظ کفایت و همگنی بررسی شده و پس از اطمینان از صحت آن‌ها، مابقی مراحل انجام و نتایج ارائه شده است.

نتایج بررسی اولیه داده‌ها از لحاظ آماری

در جدول ۱ مقادیر آماری مربوطه به داده‌های موجود آورده شده است. مواردی چون تعداد، میانگین، مینیمم، ماکزیمم و ... که برای هر پارامتر به صورت جداگانه مشخص شده است.

نتایج تکنیک Feature Selection

بر اساس اعمال تکنیک feature selection نتایج زیر حاصل شد:

با توجه به جدول فوق، سطح ایستابی با مقدار ۰/۱۴۷ مهم‌ترین فاکتور در پیش‌بینی غلظت نیترات بوده و HCO_3 و Mg به‌ترتیب با مقادیر ۰/۱۴۶ و ۰/۱۲ در رتبه‌ی دوم و سوم این جدول قرار دارند. نتایج حاصل از این تکنیک با نتایج حاصل از پژوهش بوی و همکاران کاملا مطابقت دارد (Bui et al., 2020). تنها تفاوت در وجود متغیر سطح ایستابی در پژوهش حاضر است که با مقداری بسیار نزدیک به HCO_3 در رتبه اول قرار دارد. فاکتور پتاسیم در میزان نیترات پیش‌بینی شده نقشی نداشته و بنابراین برای مدل‌سازی از ورودی‌های مدل حذف شد.

جدول ۱- خروجی بررسی آماری داده‌ها

	T	Cl	Water Level	SO_4	HCO_3	pH	TDS	Th	Ec	Na%	K	Mg	Ca	NO_3
Count	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸	۱۲۳۸
Mean	۱۸/۶۴۵	۷/۲۰۳	۲۲/۵۵۰	۲/۴۵۹	۷/۷۳۰	۷/۵۹۸	۱۱۶۹/۰۷	۵۰۱/۹	۱۷۶۳/۷	۳۲/۹۲	۰/۱۰۵	۴/۱۱۵	۵/۹۲	۹/۱۴
Std	۵/۰۹۰۸	۱۲/۴۰۹	۴۸۴/۴۲	۳/۶۳۱	۳/۰۰۸	۰/۳۶۴	۱۰۴۳/۰۸	۳۷۰/۰۹	۱۵۴۸/۶	۱۹/۸۷۷	۰/۰۲۵	۵/۰۸۱	۴/۱	۲۰/۸
Min	۳/۱	۰/۱	۰/۱۵	۰/۱	۰/۵	۶/۵	۱۲۵	۲۰	۱۹۸	۰/۷۵	۰/۰۱	۰/۱	۰/۳	۰
25%	۱۶	۱/۶	۲/۰۷۲	۰/۶	۵/۹	۷/۳	۶۵۷/۲۵	۳۴۵	۱۰۱۲/۷۵	۱۷/۷۶۲	۰/۰۹۵	۱/۸	۱,۴	۰
50%	۱۹	۳/۸	۵/۱۵۵	۱/۴	۷/۱	۷/۶	۸۷۸	۴۳۰	۱۳۳۰	۲۹/۱۶۵	۰/۱۱	۳/۳	۵/۳	۰
75%	۲۱/۱۷۵	۶/۹	۱۳/۸۱	۲/۸	۹/۱	۷/۸	۱۲۲۳/۵	۵۵۰	۱۸۲۶/۵	۴۵/۹۶۵	۰/۱۱	۴/۸	۶/۸	۹/۷
Max	۱۲۱	۱۰۹	۱۷۰۵۱	۳۹/۷	۳۰/۶	۹/۳	۹۴۸۰	۶۴۴۰	۱۴۱۵۰	۹۱/۳۳	۰/۲۱	۱۲۷/۷	۸۲۲	۲۲۵/۷

جدول ۲- خروجی تکنیک feature selection

No.	Feature	Feature importance
۱	Water level	**۰/۱۴۷**
۲	HCO_3	**۰/۱۴۶**
۳	Mg	**۰/۱۲**
۴	pH	**۰/۱۰۶**
۵	Na%	**۰/۰۹۳**
۶	SO_4	**۰/۰۹۲**
۷	T	**۰/۰۸۱**
۸	Cl	**۰/۰۷۷**
۹	Th	**۰/۰۵۵**
۱۰	Ca	**۰/۰۳۲**
۱۱	TDS	**۰/۰۳۰**
۱۲	EC	**۰/۰۱۴**
۱۳	K	**۰**

جدول ۳- ضرایب متغیرهای ورودی (شیب معادله)

T	Cl	Water Level	SO_4	HCO_3	Ph	TDS	Ec	Th	Na%	Mg	Ca
-۰/۰۰۱۶	۰/۰۰۴۹	-۰/۰۰۶۳	-۰/۰۰۳۰۵	-۰/۰۰۱۸	-۰/۰۰۱۴	-۰/۰۰۰۳	۰/۰۰۲۷	۰/۰۰۵	-۰/۰۰۰۶	۰/۰۰۸	-۰/۰۰۸

نتایج مدل رگرسیون لجستیک

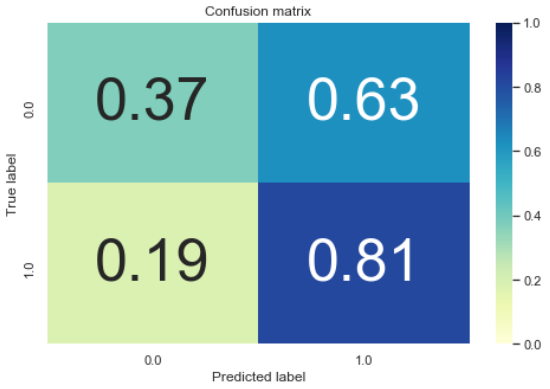
با توجه به اعمال مدل رگرسیون لجستیک روی داده‌های موجود، نتایج زیر حاصل شد. طبق تابع رگرسیون لجستیک، مقادیر شیب خط مربوط به هر متغیر در جدول زیر آورده شده است. همچنین، عرض از مبدا نیز برابر ۰/۰۰۰۳۶۳- شده است. پس از اعمال روش‌های ارزیابی مذکور، نتایج جدول ۴ به‌دست آمده است.

جدول ٤- نتایج شاخص‌های ارزیابی مدل رگرسیون لجستیک

R^2-score	NSE	Pbias	RMSE
۰/۰۴۴۲۱۰۶۷۷	-۰/۶۲۲۵۹۶۱۵۴	۴/۵۸۰۱۹	۰/۶۹۹

یکی دیگر از شاخص‌های ارزیابی استفاده شده confusion matrix می‌باشد که به خوبی نمایانگر میزان صحت کارکرد مدل است (شکل ۵). این نمودار از دو محور مقادیر واقعی و مقادیر پیش‌بینی شده متشکل است. تقاطع اعداد صفر باهم و تقاطع اعداد یک باهم نمایانگر مقادیر پیش‌بینی شده‌ی صحیح هستند. بدین معنی که پیش‌بینی‌های انجام شده توسط مدل با مقادیر واقعی همپوشانی داشته است. در طرف مقابل تقاطع اعداد صفر و یک باهم نمایانگر پیش‌بینی‌های اشتباه مدل هستند. بدین معنی که فرضا مدل پیش‌بینی می‌کند که آلودگی نیترات وجود ندارد در حالی که در داده‌های واقعی آلودگی نیترات وجود داشته است. همان‌طور که در

نمودار مشاهده می‌گردد، مدل در ۳۷ درصد موارد، پیش‌بینی عدم آلودگی صحیح و در ۸۱ درصد موارد پیش‌بینی آلودگی صحیح داشته است. در طرف مقابل، مدل در ۱۹ درصد موارد نیز پیش‌بینی عدم آلودگی اشتباه داشته که در حقیقت در داده‌های اصلی آلودگی مشهود بوده است. همچنین، مدل در ۶۳ درصد موارد نیز پیش‌بینی آلودگی اشتباه داشته است. بدین معنی که مدل پیش‌بینی کرده آب آلوده به نیترات می‌باشد در صورتی که در حقیقت آلودگی نیترات وجود نداشته است.



شکل ۵- خروجی شاخص confusion matrix مدل رگرسیون لجستیک

نتایج مدل درخت تصمیم

با توجه به اعمال مدل درخت تصمیم روی داده‌های موجود، نتایج زیر حاصل شد.

جدول ۵- نتایج شاخص‌های اصلی ارزیابی مدل درخت تصمیم

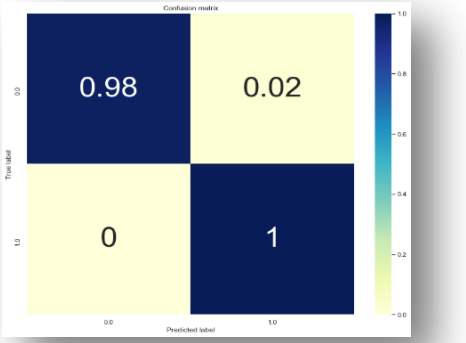
R^2-score	NSE	Pbias	RMSE
۰/۹۵۷	۰/۹۵۶	۰/۳۶۷	۰/۲۹۷

مدل آموزش داده شده‌ی درخت تصمیم:

DecisionTreeClassifier (max_depth=20, min_samples_leaf=5)

نتیجه‌ی شاخص confusion matrix:

همان‌طور که در شکل ۶ مشهود است، مدل در ۹۸ درصد موارد پیش‌بینی عدم آلودگی صحیح و در ۱۰۰ درصد موارد پیش‌بینی آلودگی صحیح داشته است. تنها در ۲ درصد موارد پیش‌بینی آلودگی اشتباه داشته است. مدل پیش‌بینی عدم آلودگی اشتباه نداشته که یکی از نقاط قوت این مدل می‌باشد.



شکل ٦- خروجی شاخص confusion matrix مدل درخت تصمیم

طبق نتایج جدول ۵، مدل درخت تصمیم با مقادیر R^2 ۰/۹۵۷ = و RMSE=۰/۲۹۷ و همچنین، با توجه به مقادیر نمودار confusion matrix از عملکرد بسیار مطلوبی برخوردار بوده است. مقدار شاخص NSE نیز نشان دهنده‌ی قدرت مدل درخت تصمیم در پیش‌بینی نیترات می‌باشد 1, Thomson et al., ≥0.75 NSE (2007). نتایج این قسمت از پژوهش با یافته‌های رودریگز و همکاران و بوی و همکاران که در آن از ورودی‌های مشابه پژوهش حاضر استفاده شد همخوانی دارد (Rodriguez-Galiano et al., 2014; Bui et al., 2020). آن‌ها نشان دادند که مدل درخت تصمیم در پیش‌بینی نیترات عملکرد بسیار خوبی نسبت به مدل رگرسیون لجستیک داشت. در ادامه، تصویر قسمتی از مدل درخت تصمیم به دست آمده، آورده شده است.

نتایج مدل شبکه عصبی مصنوعی

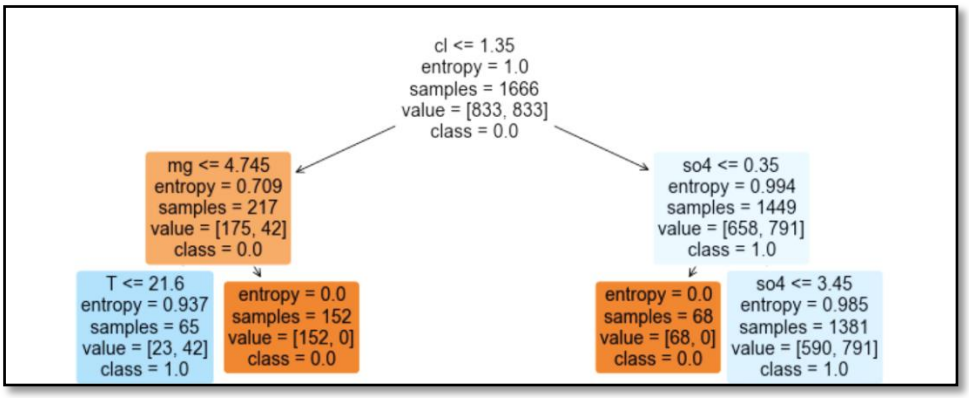
نتایج شاخص‌های اصلی ارزیابی مدل شبکه عصبی مصنوعی در جدول ۶ آورده شده است.

جدول ٦- نتایج شاخص‌های ارزیابی مدل شبکه عصبی مصنوعی

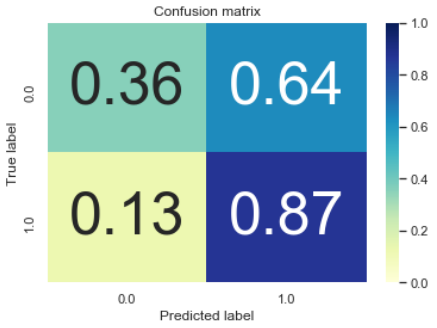
R^2-score	NSE	Pbias	RMSE (W/m^2)
۰/۰۱	−۰/۷۵	۱/۸۷	۰/۵۲

نتیجه شاخص confusion matrix

با توجه به شکل ۸ مدل شبکه عصبی مصنوعی در ۸۷ درصد موارد پیش‌بینی آلودگی صحیح و در ۶۴ درصد موارد پیش‌بینی آلودگی اشتباه داشته است. همچنین مدل در ۳۶ درصد موارد پیش‌بینی عدم آلودگی صحیح و در ۱۳ درصد موارد پیش‌بینی عدم آلودگی اشتباه داشته است.



شکل ۷- قسمتی از درخت تصمیم



شکل ۸- خروجی شاخص confusion matrix مدل شبکه عصبی مصنوعی

طبق نتایج مذکور در جدول۶ مدل شبکه عصبی مصنوعی با وجود تکنیک hyperparameter tuning با مقادیر R^2 = ۰/۰۱ و RMSE=۰/۵۲ و همچنین با توجه به مقادیر نمودار confusion matrix از عملکرد بسیار ضعیفی برخوردار بوده است. نتایج این قسمت از پژوهش با یافته‌های نولان و همکاران مطابقت دارد (Nolan et al., 2015). طبق مطالعه‌ی این محققین در پیش‌بینی نیترات آب زیرزمینی در کالیفرنیای آمریکا، مدل شبکه عصبی مصنوعی از عملکرد ضعیف‌تری نسبت به سایر مدل‌ها برخوردار بوده است.

مقایسه نتایج مدل‌ها

نتایج شاخص‌های ارزیابی مدل‌ها در جدول ۷ ارائه شده است.

جدول ۷- نتایج شاخص‌های ارزیابی مدل‌ها

Model	R^2-score	NSE	P_{bias}	RMSE
ANN	۰/۰۱	-۰/۷۵	۱/۸۷	۰/۵۲
DT	۰/۹۵	۰/۹۵	۰/۳۶	۰/۲۹
LR	۰/۰۴	-۰/۶۲	۴/۸۵	۰/۶۹

با توجه به نتایج به‌دست آمده از شاخص‌های ارزیابی برای ناحیه مورد نظر و بالاخص شاخص R^2، مدل درخت تصمیم از بقیه مدل‌ها عملکرد بسیار بهتری داشته است. پس از آن مدل رگرسیون لجستیک و در نهایت مدل شبکه عصبی مصنوعی که ضعیف‌ترین عملکرد را داشته است. نتایج این پژوهش با یافته‌های ادامری و همکاران و هی و همکاران مطابقت دارد (He et al., 2022; Uddameri et al., 2020). در مطالعات نام برده شده نیز مدل درخت تصمیم نسبت به مدل‌های دیگر عملکرد بهتری داشته است.

نتیجه‌گیری

وجود نیترات در آب‌های زیرزمینی از دو جنبه شرب و نیز آبیاری زمین‌های کشاورزی قابل بررسی است. در این پژوهش کارآمدی از الگوریتم‌های محاسباتی نرم برای پیش‌بینی نیترات آب‌های زیرزمینی محدوده شرق مازندران استفاده شد. متغیرهای سطح ایستابی و HCO_3 به ترتیب با مقادیر ۰/۱۴۷ و ۰/۱۴۶ بیشترین اهمیت را در پیش‌بینی غلظت نیترات داشتند. متغیر پتاسیم تاثیری در پیش‌بینی غلظت نیترات نداشت. مدل رگرسیون لجستیک و شبکه عصبی مصنوعی به ترتیب با مقادیر ضریب تعیین ۰/۰۴ و ۰/۰۱ عملکرد مطلوبی در پیش‌بینی غلظت نیترات نداشتند و مدل درخت تصمیم با ضریب تعیین ۰/۹۵ عملکرد بسیار مطلوبی در پیش‌بینی غلظت نیترات داشت. نتایج این پژوهش با یافته‌های دیگر محققان از جمله نولان و همکاران، ادامری و همکاران و هی و همکاران مطابقت داشت (He et al., 2022; Uddameri et al., 2020; Nolan et al., 2015). پیشنهاد می‌شود که در پژوهش‌های آتی، متغیرهایی مانند کاربری اراضی و هدایت هیدرولیکی به موارد متغیرهای ورودی اضافه شده و نتایج با پژوهش حاضر مقایسه گردد.

منابع

حسینی وردنجانی، س. م. ر.، خوش روش، م.، طاهری سودجانی، ه.، قهرمان شهرکی، م. و پورغلام آمیجی، م. ۱۴۰۲. ارزیابی کیفی آب زیرزمینی برای مصارف شرب بر اساس شاخص‌های کیفیت آب. مهندسی آبیاری و آب ایران. ۱۴ (۲): ۱۸۰-۱۶۴.

Awais, M., Aslam, B., Maqsoom, A., Khalil, U., Ullah, F., Azam, S. and Imran, M. 2021. Assessing Nitrate Contamination Risks in Groundwater: A Machine Learning Approach. Applied Sciences. 11: 10034.

Band, B. Sh., Janizadeh, S., Chandra Pal, S., Chowdhuri, I., Siabi, Zh., Norouzi, A., M. Melesse, A., Shokri, M. and Mosavi, A. 2020. Comparative Analysis of Artificial Intelligence Models for Accurate Estimation of Groundwater Nitrate Concentration, Sensors. 20: 5763.

Bui, D.T., Khosravi, Kh., Karimi, M., Busico, G., Sheikh Khozani, Z., Nguyen, H., Mastrocicco, M., Tedesco, D., Cuoco, E. and Kazakis, N. 2020. Science of the Total Environment. 715136836

Elzain, H.E., Sang Yong Chung, S.Y., Senapathi, V., Sekar, S., Lee, S.Y., Priyadarsi D., Roy, Amjed Hassan, A. and Sabarathinam, Ch. 2022. Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination. Ecotoxicology and Environmental Safety. 229: 113061

Gangolli, S.D., Van den Brandt, P.A., Feron, V.J., Janzowsky, C., Koeman, J.H., Speijers, G.J.A., Spiegelhalder, B., Walker, R. and Wishnok, J.S. 1994. Assessment of nitrate, nitrite and N-nitroso compounds: Eur. J. Pharmacol. Environ. Toxicol. Pharmacol. Section. 292: 1–38.

García-del-Toro, E.M., García-Salgado, S., Mateo, L.F., Quijano, M.Á. and Más-López, M.I. 2022. Machine Learning as a Diagnosis Tool of Groundwater Quality in Zones with High Agricultural Activity (Region of Campo de Cartagena, Murcia, Spain). Agronomy. 12: 3076.

Gholami, V. and Booij, M.J. 2022. Use of machine learning and geographical information system to predict nitrate concentration in an unconfined aquifer in Iran. Journal of Cleaner Production, 360, 131847.

He, S., Jianhua Wu, J., Dan Wang, D. and He, X. 2022. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere. 290: 133388.

Hosseini, S.M. and Mahjouri, N. 2014. Developing a fuzzy neural network-based support vector regression (FNN-SVR) for regionalizing nitrate concentration in groundwater. Environmental Monitoring and Assessment. 186: 3685–3699.

Ijlil, S., Essahlaoui, A., Mohajane, M., Essahlaoui, N., Mili, E.M. and Rompaey, V. 2022. A machine learning algorithm for modeling and mapping of groundwater pollution risk: A study to reach water security and sustainable development (Sdg) goals in a Mediterranean aquifer system. Remote Sensing. 14, 2379.

Iranian Ministry of Energy (IMOF). 2014. Rehabilitation and Balance Program for Groundwater Resources (106 pp).

Ma, L., Hu, L., Feng, X. and Songlin Wang, S. 2018. Nitrate and Nitrite in Health and Disease. Aging and disease. 9(5): 938-945.

Neshat, A., Pradhan, B., Pirasteh, S. and Shafri, H.Z.M. 2014. Estimating groundwater vulnerability to pollution using a modified DRASTIC model in the Kerman agricultural area, Iran. Environmental Earth Sciences. 71(7): 3119–3131.

Nolan, B.T., Gronberg, J.M., Faunt, C.C., Eberts, S.M. and Belitz, K. 2014. Modeling nitrate at domestic and public-supply well depths in the Central Valley, California. Environmental Science & Technology. 48: 5643–5651.

Rodriguez-Galiano V., Mendes, M.P., Garcia-Soldado, M.J., ChicaOlmo, M. and Ribeiro, L. 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Science of the Total Environment. 476: 189–206.

Rokhshad, A.M., Khashei Siuki, A. and Yaghoobzadeh, M. 2021. Evaluation of a machine-based learning method to estimate the rate of nitrate penetration and groundwater contamination, Arabian Journal of Geosciences. 14: 40.

Sajedi-Hosseini. F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F. and Pradhan, B. 2018. Science of the Total Environment. 644: 954–962.

Thomson, B.M., Nokes, C.J. and Cressey, P.J. 2007. Intake and risk assessment of nitrate and nitrite from New Zealand foods and drinking water. Fd Addit. Contam. 24: 113–121.

Uddameri, V., Bessa Silva, A.L., Singaraju, S., Mohammadi, Gh. and Hernandez, E.A. 2020. Tree-Based Modeling Methods to Predict Nitrate Exceedances in the Ogallala Aquifer in Texas. Water. 12: 1023.

WHO. 1995. Evaluation of certain food additives and contaminants. 44th report of the Joint FAO/WHO Expert Committee on Food Additives. Technical Report Series. 859: 29–35.

Iranian Journal of Irrigation and Drainage
No.2, Vol. 18, Jun.-Jul. 2024, p. 307-317

نشریه آبیاری و زهکشی ایران
شماره۲، جلد ۱۸، خرداد - تیر ۱۴۰۳، ص. ۳۱۷-۳۰۷

Prediction of Nitrate Concentration in Groundwater of the Eastern Region of Mazandaran Province using Soft Computing Algorithms

F. Doustalizadeh[1], M. Khoshravesh[2*], R. R. Fazloula[3], M.M. Bateni[4]
Recived: Dec.01, 2023 Accepted: Jun.06, 2024

Abstract

Considering the importance of fresh water for human life and the vulnerability of groundwater sources to all kinds of pollution and the possibility of transferring pollutants to other surface and groundwater sources, as well as the location of Iran in the arid and semi-arid belt, protecting this valuable and rare element is imperative and its continuous monitoring must be one of the priorities of water resources managers. Therefore, in the current research, nitrate pollution in the eastern plains of Mazandaran province was discussed and relevant issues were investigated, and an efficient and optimal model for predicting nitrate concentration was presented. In this research, three machine learning models including decision tree, logistic regression and artificial neural network were compared. The physical and chemical data measured during the years 1985 to 2020 were used and entered as the input variables of the models. The variables include temperature, water level, pH, EC, HCO_3^-, CL^-, SO_2^{4-}, Na^+, K^+, Mg^{2+}, Ca^{2+}, TH and TDS; The amount of nitrate contamination of the groundwater, was predicted by dividing 70% of the dataset as training and 30% as testing data. The R^2, RMSE, NSE and PBIAS indexes were applied for model evaluation. The results indicated that the Decision Tree model had the best performance with a large difference compared to the other two models (R^2 = 0.957 and RMSE = 0.297, NSE = 0.95 and testing acc = 0.907). After That, logistic regression and artificial neural network had much weaker performances than the lead model. It is suggested to conduct another research with other machine learning models by changing the input variables and add some extra ones such as land-use and compare the results with the current research.

Keywords: Artificial neural network, Data mining, Decision trees, Logistic regression

1- MSc. Graduated of Irrigation and Drainage, Water Engineering Department, Faculty of Agricultural Engineering, Sari Agricultural Sciences and Natural Resources University, Sari, Iran
2- Associate Professor, Water Engineering Department, Faculty of Agricultural Engineering, Sari Agricultural Sciences and Natural Resources University, Sari, Iran
3- Associate Professor, Department of Water Science and Engineering, Faculty of Agricultural Engineering, Sari Agricultural Sciences and Natural Resources University, Sari, Iran
4- Researcher, Scuola Superiore Studi Pavia IUSS, Pavia, Italy
(*- Corresponding Author: m.khoshravesh@sanru.ac.ir)