Vol 16, No. 3, Autumn 2026

University of Tehran Press

Journal of Water and Irrigation Management Online ISSN: 2382-9931

Homepage: https://jwim.ut.ac.ir/

Spatially Informed Machine-Learning Prediction of Groundwater Quality Using Feature Engineering from Proximal Observation Wells

Hossein Hadizadeh[1] | Nima Moghaddam[2] | Afshin Ashrafzadeh[3]✉

1. Department of Irrigation and Reclamation Engineering, Faculty of Agriculture and Natural Resources, University of Tehran, Karaj, Iran. E-mail: hossein.hadizadeh@ut.ac.ir
2. Department of Irrigation and Reclamation Engineering, Faculty of Agriculture and Natural Resources, University of Tehran, Karaj, Iran. E-mail: nmoghaddam@ut.ac.ir
3. Corresponding Author, Department of Irrigation and Reclamation Engineering, Faculty of Agriculture and Natural Resources, University of Tehran, Karaj, Iran. E-mail: A.Ashrafzadeh@ut.ac.ir

Article Info

Article type:
Research Article

Article history:
Received 6 May 2026
Received in revised form
15 June 2026
Accepted 6 July 2026
Published online 26 October 2026

Keywords:
Groundwater quality prediction
Machine learning
Spatial feature engineering
Electrical conductivity (EC)
Calcium (Ca^{2+})

ABSTRACT

Accurate modeling of groundwater quality requires a proper understanding of the spatial dependency structure of hydrochemical parameters and the application of advanced predictive approaches. In this study, six machine learning algorithms including K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), Random Forest (RF), Gaussian Process Regression (GPR), Gradient Boosting Regression (GBR) and Extreme Gradient Boosting (XGBoost) were employed to predict electrical conductivity (EC) and calcium (Ca^{2+}) concentrations in Guilan Province, Iran. To incorporate spatial dependency into the predictive framework, the distances to neighboring wells were introduced as input features, and model performance was evaluated under different neighborhood scenarios (4 to 8 neighboring wells) using annual groundwater quality data from 2002 to 2018. The results indicate that ensemble learning models outperform the other algorithms; however, the optimal spatial configuration varies between parameters. For electrical conductivity (EC), the Random Forest model achieved the best performance when eight neighboring wells were considered, with an R-squared value of 0.888, a root mean square error of 117.165, and a mean absolute error of 84.173. This suggests a broader spatial dependency for this parameter within the aquifer system. In contrast, for calcium ion (Ca^{2+}), the XGBoost model yielded the optimal performance with four neighboring wells, achieving an R-squared value of 0.811, a root mean square error of 1.154, and a mean absolute error of 0.820, indicating that local hydrogeochemical processes exert a stronger control on the spatial distribution of this ion. The comparison of neighborhood scenarios further demonstrates that increasing the number of neighboring wells does not necessarily improve prediction accuracy for all parameters, highlighting the importance of parameter-specific spatial dependency analysis. Overall, the findings confirm that integrating machine learning with spatial feature engineering provides an effective framework for intelligent groundwater quality modeling and can support the optimization of monitoring networks and water resources management strategies.

Cite this article: Hadizadeh, H., Moghadam, N., & Ashrafzadeh, A. (2026). Spatially Informed Machine-Learning Prediction of Groundwater Quality Using Feature Engineering from Proximal Observation Wells. *Journal of Water and Irrigation Management*, 16 (3), 499-519. DOI: https://doi.org/10.22059/jwim.2026.413821.1301

© The Author(s).
Publisher: University of Tehran Press.
DOI: https://doi.org/10.22059/jwim.2026.413821.1301

دوره شانزدهم، شماره سوم، پاییز ۱۴۰۵

انتشارات دانشگاه تهران

مدیریت آب و آبیاری

شاپا الکترونیکی: ۹۹۳۱-۲۳۸۲

Homepage: https://jwim.ut.ac.ir/

پیش‌بینی متغیرهای کیفی آب زیرزمینی با یادگیری ماشین و مهندسی ویژگی‌های فضایی چاه‌های همجوار

حسین هادی‌زاده[۱] | نیما مقدم[۲] | افشین اشرف‌زاده[۳]✉

۱. گروه مهندسی آبیاری و آبادانی دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: hossein.hadizadeh@ut.ac.ir

۲. گروه مهندسی آبیاری و آبادانی دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: nmoghaddam@ut.ac.ir

۳. نویسنده مسئول، گروه مهندسی آبیاری و آبادانی دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: A.Ashrafzadeh@ut.ac.ir

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۵/۰۲/۱۶

تاریخ بازنگری: ۱۴۰۵/۰۳/۲۵

تاریخ پذیرش: ۱۴۰۵/۰۴/۱۵

تاریخ انتشار: ۱۴۰۵/۰۸/۰۴

کلیدواژه‌ها:

پیش‌بینی کیفیت آب زیرزمینی

یادگیری ماشین

مهندسی ویژگی‌های فضایی

شوری

کلسیم

چکیده

مدل‌سازی دقیق کیفیت آب‌های زیرزمینی مستلزم درک ساختار وابستگی مکانی پارامترهای شیمیایی و بهره‌گیری از روش‌های پیش‌بینی پیشرفته است. در این پژوهش، با هدف پیش‌بینی متغیرهای هدایت الکتریکی (EC) و کلسیم (Ca^{+2}) در استان گیلان، از شش الگوریتم یادگیری ماشین شامل KNN، MLP، Random Forest (RF)، Gaussian Process Regression (GPR)، Gradient Boosting Regression (GBR) و XGBoost استفاده شد. به‌منظور لحاظ‌کردن ساختار فضایی، فاصله چاه‌های همجوار به‌عنوان ورودی به مدل‌ها افزوده شد و عملکرد آن‌ها در سناریوهای مختلف تعداد چاه‌های همسایه (چهار تا هشت چاه) با استفاده از داده‌های سالانه دوره ۲۰۰۲ تا ۲۰۱۸ ارزیابی گردید. نتایج نشان داد که مدل‌های مبتنی بر یادگیری جمعی عملکرد برتری نسبت به سایر الگوریتم‌ها دارند، اما الگوی بهینه فضایی برای هر متغیر متفاوت است. برای EC، مدل RF در سناریوی استفاده از هشت چاه همسایه بهترین عملکرد را با مقدار R^2 برابر با ۰/۸۸۸، RMSE برابر با ۱۱۷/۱۶۵ و MAE برابر با ۸۴/۱۷۳ ارائه داد که نشان‌دهنده وابستگی مکانی گسترده‌تر این متغیر در آبخوان است. در مقابل، برای Ca^{+2} مدل XGBoost در سناریوی چهار چاه همسایه به عملکرد بهینه با R^2 برابر با ۰/۸۱۱ و MAE برابر با ۰/۸۲۰ دست یافت که بیانگر غالب‌بودن فرایندهای محلی در توزیع مکانی این یون است. مقایسه سناریوها نشان داد که افزایش تعداد چاه‌های همسایه الزاماً منجر به بهبود دقت پیش‌بینی برای تمامی پارامترها نمی‌شود و ساختار وابستگی مکانی هر متغیر باید جداگانه تحلیل شود. یافته‌های این مطالعه تأکید می‌کند که تلفیق یادگیری ماشین با مهندسی ویژگی‌های فضایی، چارچوبی مؤثر برای مدل‌سازی هوشمند کیفیت آب‌های زیرزمینی فراهم می‌کند و می‌تواند در بهینه‌سازی شبکه‌های پایش و تصمیم‌گیری مدیریتی مورداستفاده قرار گیرد.

استناد: هادی‌زاده، حسین؛ مقدم، نیما و اشرف‌زاده، افشین (۱۴۰۵). پیش‌بینی متغیرهای کیفی آب زیرزمینی با یادگیری ماشین و مهندسی ویژگی‌های فضایی چاه‌های همجوار. *نشریه مدیریت آب و آبیاری*، ۱۶ (۳)، ۴۹۹-۵۱۹. DOI: https://doi.org/10.22059/jwim.2026.413821.1301

ناشر: مؤسسه انتشارات دانشگاه تهران.

© نویسندگان.

۱. مقدمه

کمبود منابع آب شیرین در سطح جهانی و به‌ویژه در منطقه غرب آسیا و شمال آفریقا، به یکی از تهدیدات جدی برای امنیت آبی، غذایی و اجتماعی تبدیل شده است. ایران به‌عنوان کشوری با اقلیم خشک و نیمه‌خشک، با محدودیت‌های شدید منابع آب مواجه است. براساس آمارهای رسمی، بخش کشاورزی با مصرف حدود ۹۰ درصد از کل منابع آب مصرفی کشور، اصلی‌ترین عامل فشار بر منابع آب زیرزمینی محسوب می‌شود (معاونت برنامه‌ریزی و توسعه منابع آب، ۱۴۰۰).

رشد جمعیت، گسترش اراضی زیر کشت، تغییر الگوی کشت به سمت محصولات آب‌بر و هم‌چنین تغییرات اقلیمی که منجر به کاهش بارش‌ها و افزایش دما شده است، تعادل طبیعی چرخه هیدرولوژیک را بر هم زده است. در این شرایط، آب‌های زیرزمینی که به‌عنوان ذخیره‌ای استراتژیک برای روزهای پرتلاطم خشک‌سالی عمل می‌کنند، با خطر برداشت بیش از حد روبه‌رو هستند. این برداشت‌های بی‌رویه منجر به افت شدید سطح ایستابی، کاهش ظرفیت ذخیره‌سازی، افت کیفیت آب به‌دلیل تغلیظ آلاینده‌ها و پدیده‌های خطرناک محیط‌زیستی مانند فرونشست زمین و نفوذ آب شور در دشت‌های ساحلی و داخلی شده است (Wada *et al*., 2010). بنابراین، مدیریت پایدار منابع آب زیرزمینی دیگر یک انتخاب نیست، بلکه یک ضرورت حیاتی برای بقای اکوسیستم‌ها و تداوم فعالیت‌های اقتصادی و اجتماعی کشور است. در این میان، پایش و مدل‌سازی کیفیت آب‌های زیرزمینی با دقت بالا به‌عنوان یکی از ارکان اصلی مدیریت یکپارچه منابع آب[۱] اهمیت ویژه‌ای پیدا کرده است. کیفیت آب تحت تأثیر طیف وسیعی از عوامل طبیعی و انسانی است. عوامل طبیعی شامل فرایندهای هوازدگی سنگ‌ها، تبادل یونی، انحلال املاح و واکنش‌های بیوشیمیایی در سفره‌های آب است (Gurdak & Qi, 2012; Moghaddam & Ashrafzadeh, 2026). از سوی دیگر، عوامل انسانی مانند استفاده از کودهای شیمیایی و سموم کشاورزی، ورود فاضلاب‌های شهری و صنعتی تصفیه‌نشده و فعالیت‌های معدنی، ترکیب شیمیایی آب‌های زیرزمینی را به شدت تغییر داده‌اند. آلاینده‌هایی مانند نیترات، نیتریت، فلزات سنگین (مانند آرسنیک، سرب و کادمیوم) و سختی کل، سلامت جامعه و پایداری کشاورزی را تهدید می‌کنند (Alijani *et al*., 2018; Moghaddam & Tavana, 2023).

در دهه اخیر، کاربرد الگوریتم‌های یادگیری ماشین در پیش‌بینی کیفیت آب زیرزمینی گسترش قابل‌توجهی یافته است. پژوهش‌گران متعددی نشان داده‌اند که این الگوریتم‌ها می‌توانند روابط پیچیده و غیرخطی بین پارامترهای هیدروشیمیایی، ویژگی‌های هیدروژئولوژیکی و کیفیت آب را با دقتی بالاتر از روش‌های آماری کلاسیک (مانند رگرسیون خطی) مدل‌سازی کنند. در ادامه، مروری بر مهم‌ترین مطالعات انجام‌شده در این حوزه ارائه می‌شود. برخی مطالعات بر کاربرد شبکه‌های عصبی مصنوعی[۳] (ANN) تمرکز داشته‌اند.

روش‌های سنتی پایش کیفیت آب که مبتنی بر نمونه‌برداری فیزیکی از چاه‌ها و آزمایش‌های آزمایشگاهی هستند، اگرچه دقیق‌اند، اما دارای معایب جدی شامل هزینه‌های بالا، زمان‌بربودن، پراکندگی مکانی نامناسب و عدم توانایی در پیش‌بینی تغییرات آتی هستند. درحقیقت روش‌های سنتی داده‌های لازمه را برای بهره‌گیری از روش‌های نوین تهیه می‌کنند. با توجه به پویایی و پیچیدگی سیستم‌های آب‌زیرزمینی، نیاز به روش‌های نوین، سریع و کم‌هزینه‌ای احساس می‌شود که بتوانند با استفاده از داده‌های موجود، الگوهای تغییرات کیفیت آب را شناسایی کرده و برای آینده پیش‌بینی کنند. در این راستا، رویکردهای مبتنی بر هوش مصنوعی و یادگیری ماشین[۴] به‌عنوان ابزارهایی قدرتمند برای حل مسائل پیچیده غیرخطی در هیدرولوژی ظهور کرده‌اند (Singh *et al*., 2019).

مقدم و همکاران (۲۰۲۵) توانستند با درنظرگرفتن ترکیب الگوریتم‌های مختلف هوش مصنوعی و لحاظ‌کردن و واردکردن فاصله و اطلاعات چاه‌های همسایه به‌عنوان بخشی از متغیرهای ورودی با مقایسه مکانی بخشی از حوضه

آبریز دریاچه نمک و دریای خزر دریافتند که درمجموع مدل جنگل تصادفی بهترین عملکرد را از بین مدل‌های مورداستفاده آن‌ها داشته است (Moghaddam *et al*., 2025). هم‌چنین، دو[۵] و همکاران (۲۰۲۲) در حوضه آبریز هولا[۶] در چین با استفاده از شبکه عصبی کانولوشنی (CNN) و پردازش تصویر برای تشخیص خودکار تراز آب، به دقت ۹۴/۶ درصد دست یافتند و نشان دادند که مدل CNN عملکردی به‌مراتب بهتر از روش تطبیق الگو (Template Matching) دارد.

کومار[۷] و همکاران (۲۰۲۰) از مدل‌های شبکه‌های عصبی مصنوعی[۸] برای پیش‌بینی غلظت آلاینده‌های آلی در آب‌های زیرزمینی منطقه‌ای در هند استفاده کردند و نشان دادند که این مدل‌ها در یادگیری الگوهای پیچیده عملکرد خوبی دارند، اما همواره با چالش همگرایی در مراحل آموزش و حساسیت به پارامترهای اولیه روبه‌رو هستند. در مطالعه‌ای دیگر، زِنگ[۹] و همکاران (۲۰۲۱) از الگوریتم جنگل تصادفی[۱۰] (RF) برای طبقه‌بندی کیفیت آب در حوضه آبریز یانگ‌تسه در چین استفاده کردند. نتایج آن‌ها نشان داد که RF به‌دلیل مقاومت بالا در برابر داده‌های پرت[۱۱] و عدم نیاز به نرمال‌سازی شدید داده‌ها، عملکردی پایدارتر از ANN دارد. هم‌چنین، لیو[۱۲] و همکاران (۲۰۲۲) از الگوریتم تقویت‌گر گرادیانی[۱۳] (GBR) برای پیش‌بینی غلظت نیترات در سفره‌های آب زیرزمینی استفاده کردند و برتری این روش را در مدیریت خطاهای پیش‌بینی اثبات نمودند (Kumar *et al*., 2020; Zeng *et al*., 2021; Liu *et al*., 2022).

هم‌چنین الگوریتم ایکس جی بوست (XGBoost)[۱۴] به‌دلیل سرعت بالا و قابلیت تنظیم دقیق پارامترها، در سال‌های اخیر موردتوجه قرار گرفته است. پژوهش‌گران متعددی از این الگوریتم در زمینه‌های مختلف هیدرولوژی استفاده کرده‌اند. برای نمونه، چن[۱۵] و همکاران (۲۰۲۱) با بررسی حوضه آبریز رودخانه فوه[۱۶] در چین، از مدل XGBoost برای پیش‌بینی کیفیت آب در شرایط مختلف اقلیمی استفاده کردند. نتایج پژوهش نشان داد که این مدل با دقت مناسبی قادر به شبیه‌سازی تغییرات کیفیت آب بوده و می‌تواند در تحلیل اثرات تغییر اقلیم بر منابع آب سطحی به‌طور مؤثری به‌کار گرفته شود.

۱. ۱. نوآوری و اهمیت پژوهش

نوآوری اصلی این پژوهش در ارزیابی تعمیم‌پذیری رویکرد مبتنی بر ساختار فضایی چاه‌های مشاهده‌ای برای متغیرهای مختلف کیفیت آب زیرزمینی است. درحالی‌که مطالعات پیشین بیش‌تر این رویکرد را برای یک متغیر کیفی خاص به‌کار گرفته‌اند و نتایج حاصل را بدون ارزیابی کافی به سایر شاخص‌های کیفی تعمیم داده‌اند، در این پژوهش دو متغیر کیفی با ویژگی‌های متفاوت به‌صورت هم‌زمان موردبررسی قرار گرفته‌اند. این موضوع امکان پاسخ‌گویی به این پرسش اساسی را فراهم می‌کند که مدل برتر و تعداد بهینه چاه‌های همسایه برای متغیرهای مختلف کیفیت آب چگونه خواهد بود.

علاوه بر این، پژوهش حاضر با مقایسه چندین الگوریتم یادگیری ماشین از جمله XGBoost، Random Forest و MLP، سایر مدل‌های متداول، چارچوبی جامع برای ارزیابی نقش ساختار فضایی داده‌ها در شبیه‌سازی کیفیت آب زیرزمینی ارائه می‌دهد. از این‌رو، برخلاف مطالعات پیشین که نتایج آن‌ها به یک متغیر یا یک مدل خاص محدود بوده است، یافته‌های این پژوهش می‌تواند مبنایی برای انتخاب مدل مناسب و تعیین تعداد بهینه چاه‌های همسایه در مدل‌سازی طیف وسیع‌تری از شاخص‌های کیفی آب زیرزمینی فراهم آورد.

اهمیت این پژوهش از آنجا ناشی می‌شود که تصمیم‌گیری‌های مدیریتی در زمینه پایش و حفاظت از منابع آب زیرزمینی به‌شدت به‌دقت پیش‌بینی متغیرهای کیفی وابسته است. در صورتی‌که تعداد چاه‌های همسایه یا مدل مورداستفاده بدون توجه به نوع متغیر کیفی انتخاب شود، احتمال کاهش دقت پیش‌بینی و افزایش عدم قطعیت در نتایج وجود خواهد داشت. انتخاب مدل‌ها بر مبنای مقایسه خانواده‌های مختلف یادگیری ماشین بوده است. بنابراین، نتایج این پژوهش می‌تواند به توسعه مدل‌های قابل اعتمادتر برای پایش کیفیت آب زیرزمینی و بهبود تصمیم‌گیری در مدیریت منابع آب کمک کند.

۲. روش‌شناسی پژوهش

۲. ۱. داده‌ها و پیش‌پردازش داده‌ها

در این پژوهش، برای مدل‌سازی کیفیت آب‌های زیرزمینی، از یک مجموعه داده شامل مختصات و مشاهدات واقعی مقادیر اندازه‌گیری‌شده از چاه‌های کیفی (متغیرهای شوری و کلسیم) برای سال‌های ۲۰۰۲ تا ۲۰۱۸ استفاده شد. اطلاعات این پژوهش از شرکت مدیریت منابع آب ایران دریافت شده است. هم‌چنین داده‌های ورودی پیش از آموزش مدل‌ها با استفاده از StandardScaler استانداردسازی شدند.

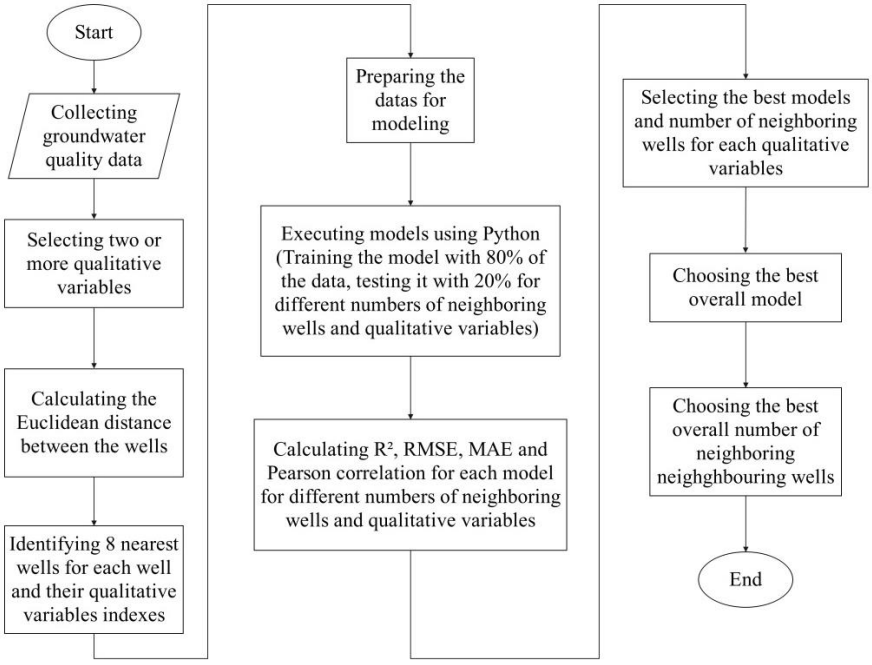


Figure 1. Flowchart of the Research Method

۲. ۲. متغیرهای ورودی و هدف

دو متغیر کیفی آب زیرزمینی شامل هدایت الکتریکی و غلظت یون کلسیم برای چاه‌های موردبررسی در این پژوهش به‌عنوان متغیرهای هدف در نظر گرفته شده و برای هر یک از آن‌ها مدل‌های پیش‌بینی به‌صورت مستقل توسعه یافت. به‌منظور لحاظ‌کردن ساختار وابستگی فضایی آبخوان، فاصله چاه‌های همسایه از چاه هدف به‌همراه مقادیر متغیر کیفی موردبررسی مربوط به آن‌ها، هدایت الکتریکی و یا غلظت یون کلسیم، به‌عنوان ویژگی‌های ورودی به مدل‌ها افزوده شد. برای ارزیابی تأثیر اطلاعات مکانی بر دقت پیش‌بینی نیز، پنج سناریوی مختلف شامل استفاده از چهار، پنج، شش، هفت و هشت چاه همسایه تعریف گردید و عملکرد مدل‌ها در هر یک از این سناریوها و برای هریک از متغبرهای کیفی مورد مقایسه قرار گرفت.

۲. ۳. الگوریتم‌های یادگیری ماشین مورداستفاده

انتخاب این شش الگوریتم با هدف پوشش دادن طیف کامل رویکردهای یادگیری ماشین انجام شد؛ به این معنا که

MLP نماینده شبکه‌های عصبی و مدل‌سازی روابط غیرخطی پیچیده، GPR نماینده مدل‌های غیرپارامتری و احتمالاتی، KNN نماینده روش‌های مبتنی بر فاصله و همسایگی، RF نماینده روش‌های تجمیعی مبتنی بر bagging و GBR و XGBoost نماینده خانواده boosting بودند. بنابراین، مقایسه هم‌زمان آن‌ها امکان ارزیابی اثر ساختار فضایی و تفاوت خانواده‌های مختلف مدل را فراهم می‌کند (Breiman, 2001; Friedman, 2001; Chen & Guestrin, 2016; Singh *et al*., 2019; Agheli *et al*., 2024; Kallab *et al*., 2024). این انتخاب از نظر کاربردی نیز موجه است، زیرا در پژوهش‌های پیشین مرتبط با مسائل آب زیرزمینی و منابع آب، شبکه‌های عصبی برای پیش‌بینی کیفیت آب زیرزمینی به‌کار رفته‌اند. برای مثال جنگل تصادفی برای ارزیابی کیفیت آب زیرزمینی، گرادیان بوستینگ برای پیش‌بینی غلظت نیترات در آب زیرزمینی و XGBoost نیز در پیش‌بینی کیفیت شاخص کیفیت آب زیرزمینی استفاده شده است (Zhang *et al*., 2021; Kumar *et al*., 2020; Liu *et al*., 2022; Chen *et al*., 2021; Subudhi *et al*., 2025). هم‌چنین، رویکردهای داده‌محور و ensemble در مدل‌سازی شوری و سایر شاخص‌های آب زیرزمینی کارآمد گزارش شده‌اند (Moghaddam *et al*., 2025; Agheli *et al*., 2024). الگوریتم‌های استفاده‌شده به شرح زیر پیاده‌سازی و تنظیم پارامترها گردیدند.

۲. ۳. ۱. شبکه‌های عصبی مصنوعی (MLP)

از مدل پرسپکترون چندلایه[۱۷] به‌عنوان نماینده روش‌های مبتنی بر شبکه‌های عصبی استفاده شد. ساختار شبکه شامل دو لایه پنهان با تعداد نودهای (۱۰۰، ۵۰) بود. این معماری براساس جستجوی سیستماتیک (Systematic exploration) و به‌منظور ایجاد تعادل بهینه بین پیچیدگی مدل و جلوگیری از بیش‌برازش انتخاب گردید. تابع فعال‌سازی در لایه‌های پنهان رِلو[۱۸] و در لایه خروجی خطی در نظر گرفته شد. برای جلوگیری از بیش‌برازش[۱۹] و برای اطمینان از تعمیم‌پذیری مدل، از روش اعتبارسنجی متقابل (Cross-Validation) استفاده شد و مقدار آستانه منظم‌سازی، مقدار آستانه منظم‌سازی (Alpha) L2 برابر با ۰/۰۰۱ (که در مطالعات مشابه هیدروژئوشیمیایی به‌عنوان مقداری بهینه تأیید شده است) و حداکثر تعداد تکرارهای آموزش برابر با ۱۰۰۰۰ (برای تضمین همگرایی کامل مدل) در نظر گرفته شد (Agheli *et al*., 2024; Kallab, 2024).

۲. ۳. ۲. همسایگان نزدیک[۲۰] (KNN)

الگوریتم KNN به‌عنوان یک روش یادگیری مبتنی بر نمونه (Instance-based learning) مورداستفاده قرار گرفت. در این پژوهش، مقدار بهینه برای تعداد همسایگان (K) با استفاده از روش اعتبارسنجی متقابل k–لایه (k-fold Cross-Validation) و با هدف کمینه‌سازی خطای پیش‌بینی تعیین گردید. معیار فاصله اقلیدسی به‌دلیل ماهیت پیوسته داده‌های هیدروژئوشیمیایی و عملکرد اثبات‌شده آن در فضای ویژگی‌های استاندارد انتخاب شد (Hastie *et al*., 2009; Agheli *et al*., 2024).

۲. ۳. ۳. جنگل تصادفی (RF)

الگوریتم RF به‌عنوان یک روش یادگیری جمعی[۲۱] مبتنی بر درخت تصمیم و روش بوت‌استراپ پیاده‌سازی شد. در این پژوهش، تعداد درخت‌های تصمیم[۲۲] برابر با ۱۰۰ در نظر گرفته شد. این مقدار براساس پایش خطای Out-of-Bag (OOB) و اعتبارسنجی متقابل انتخاب گردید، به‌طوربه‌طوری‌که مشاهده شد پس از ۱۰۰ درخت، خطای مدل به ثبات (Stabilization) رسیده و افزایش بیش‌تر درختان تأثیر معناداری بر بهبود عملکرد ندارد، ضمن این‌که اصل سادگی مدل (Model Parsimony) رعایت شده، این الگوریتم با ایجاد تنوع بین درختان از طریق نمونه‌گیری تصادفی داده‌ها و ویژگی‌ها، از خطر بیش‌برازش جلوگیری کرده و پایداری بالایی در پیش‌بینی نشان می‌دهد (Breiman, 2001; Agheli *et al*., 2024).

۲. ۳. ۴. تقویت گرادیان (GBR)

مدل GBR نیز از خانواده الگوریتم‌های تقویتی[۲۳] است که در آن مدل‌های ضعیف (معمولاً درخت‌های تصمیم) به‌صورت ترتیبی آموزش می‌بینند تا خطای مدل قبلی را جبران کنند. در این مطالعه، تعداد درختان[۲۴] برابر با ۱۰۰ این مقدار براساس ارزیابی معیارهای خطا در فرایند اعتبارسنجی انتخاب گردید تا نقطه تعادل مناسبی بین دقت مدل و هزینه محاسباتی فراهم کند و از نوسانات شدید در فرایند بهینه‌سازی جلوگیری شود (Friedman, 2001; Agheli *et al*., 2024).

۲. ۳. ۵. ایکس‌جی‌بوست (XGBoost)

الگوریتم XGBoost به‌عنوان یکی از کارآمدترین روش‌های یادگیری تقویتی پیاده‌سازی گردید. هدف از بهینه‌سازی[۲۵] بر روی خطای مجذور میانگین[۲۶] تنظیم شد. تعداد درختان برابر با ۱۰۰ و نرخ یادگیری برابر با ۰/۱ در نظر گرفته شد. این ترکیب از هایپرپارامترها براساس ارزیابی سیستماتیک عملکرد مدل انتخاب شد، زیرا تعادل بسیار خوبی بین سرعت همگرایی و جلوگیری از بیش‌برازش ایجاد می‌کند و در بسیاری از کاربردهای هیدروژئوشیمیایی عملکرد برتری داشته است (Su *et al*., 2023; Agheli *et al*., 2024).

۲. ۳. ۶. فرایند گاوسی[۲۷] (GPR)

رگرسیون فرایند گاوسی (GPR) یک رویکرد یادگیری ماشین غیرپارامتری مبتنی بر احتمالات است که متغیر هدف را به‌عنوان یک توزیع بر روی توابع مدل‌سازی می‌کند. این روش به‌ویژه برای مدل‌سازی کیفیت آب زیرزمینی به‌دلیل توانایی‌اش در ارائه پیش‌بینی‌های نقطه‌ای همراه با کمّی‌سازی دقیق عدم قطعیت، که برای ارزیابی ریسک ضروری است، مؤثر می‌باشد. در این پژوهش، از هسته RBF استفاده شد، زیرا براساس مقایسه اولیه توابع هسته، بهترین برازش را برای مدل‌سازی همبستگی‌های فضایی نرم بین چاه‌های همسایه نشان داد. هم‌چنین، هایپرپارامتر مقیاس طول (length-scale) با استفاده از روش بیشینه‌سازی درست‌نمایی (Maximum Likelihood Estimation) در یک محدوده تعریف‌شده بهینه‌سازی شد تا GPR بتواند به‌طور مؤثر با تغییرات فضایی محلی در آبخوان سازگار شود (Dalmasso *et al*., 2025; Agheli *et al*., 2024).

۲. ۴. اعتبارسنجی و ارزیابی مدل

برای ارزیابی عملکرد و پایداری تمامی الگوریتم‌های پیشنهادی، از روش اعتبارسنجی متقاطع[۲۸] (k-fold) استفاده شد. در این روش، مجموعه داده به‌صورت تصادفی به پنج بخش مساوی تقسیم گردید. در هر تکرار، چهار بخش برای آموزش[۲۹] و یک بخش برای آزمون[۳۰] استفاده شد. این فرایند پنج بار تکرار شد تا هر بخش دقیقاً یک بار در نقش مجموعه آزمون قرار گیرد. سپس میانگین معیارهای ارزیابی در بخش‌های آزمون محاسبه، گزارش و بررسی شد.

۲. ۴. ۱. معیارهای ارزیابی عملکرد

برای سنجش دقت مدل‌ها در مقایسه با مقادیر واقعی، از چهار معیار آماری رایج زیر استفاده شد:

۱. خطای مجذور میانگین (RMSE):

رابطه ۱)
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

۲. میانگین خطای مطلق (MAE):

$$MAE = \frac{1}{n}\sum_{L=1}^{n}|f_i - y_i|$$ رابطه ۲)

۳. ضریب تعیین (R-squared):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Yi - \hat{Y}_i)^2}{(Yi - \bar{y}_i)^2}$$ رابطه ۳)

۴. ضریب همبستگی پیرسون (Pearson Correlation Coefficient):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$ رابطه ۴)

۲. ۵. تحلیل حساسیت و انتخاب بهترین ویژگی‌های فضایی

برای تعیین بهترین مدل در پیش‌بینی کیفیت آب، یک تحلیل حساسیت[۳۱] انجام شد. در این بخش، میانگین عملکرد مدل‌ها در تعداد چاه‌های همجوار متفاوت، از چهار تا هشت، در نظر گرفته شد. معیار انتخاب بهترین مدل، دست‌یابی به کم‌ترین مقدار RMSE و بالاترین مقدار R^2 در مجموعه داده‌های آزمون بود. شاخص‌های Pearson correlation و MAE نیز برای ارائه تصویری کامل‌تر از جزئیات عملکرد مدل‌ها محاسبه شدند. هم‌چنین در مواردی که دو مدل از نظر R^2 و RMSE عملکرد یکسانی داشتند، برای تصمیم‌گیری نهایی از Pearson correlation و MAE استفاده شد.

۲. ۶. پیاده‌سازی نرم‌افزاری

تمامی مراحل پیش‌پردازش داده‌ها، آموزش مدل‌ها، اعتبارسنجی و تحلیل نتایج با استفاده از زبان برنامه‌نویسی پایتون (Python 3.x) و کتابخانه‌های تخصصی Scikit-learn، XGBoost، NumPy و Pandas پیاده‌سازی گردید. برای رسم نمودارهای مقایسه‌ای (مانند نمودار مقادیر واقعی در برابر پیش‌بینی شده) از کتابخانه Matplotlib استفاده شد.

۲. ۷. منطقه مطالعاتی

استان گیلان، واقع در شمال ایران و در حاشیه دریای کاسپین، به‌عنوان یکی از مهم‌ترین مناطق کشاورزی و زیست‌محیطی کشور شناخته می‌شود. این استان از نظر جغرافیایی در مختصات عرض جغرافیایی ۴۸ درجه و ۳۲ دقیقه تا ۵۰ درجه و ۳۶ دقیقه شرقی و طول جغرافیایی ۳۶ درجه و ۳۳ دقیقه تا ۳۸ درجه و ۲۷ دقیقه شمالی قرار دارد و مساحتی بالغ بر ۱۳۸۱۰/۵ کیلومترمربع را پوشش می‌دهد (Dinpashoh *et al*., 2018; Behzadi *et al*., 2021; Moghaddam *et al*., 2025). از نظر مرزی، گیلان از جنوب با استان‌های قزوین و زنجان، از شرق با مازندران و از غرب با اردبیل هم‌مرز است (شکل ۲). ساختار توپوگرافی و اقلیمی استان بسیار متنوع است. اگرچه غالب منطقه دارای آب‌وهوای معتدل و مرطوب خزری است، اما در نواحی جنوبی و جنوب‌شرقی، تأثیرات اقلیم نیمه‌خشک نیز مشاهده می‌شود. میانگین بارش سالانه در این منطقه حدود ۱۵۰۰ میلی‌متر است که نوسانات فصلی قابل‌توجهی دارد (Dinpashoh *et al*., 2018; Behzadi *et al*., 2021).

دمای میانگین ماهانه در گرم‌ترین ماه سال به ۲۹/۵ درجه سانتی‌گراد و در سردترین ماه به ۶/۳ درجه سانتی‌گراد می‌رسد. هیدرولوژی و کشاورزی استان گیلان در سطح ملی پررنگ است. این منطقه دارای حدود ۴۰۰،۰۰۰ هکتار اراضی کشاورزی است و سالانه ۷۰۰،۰۰۰ تن برنج سفید تولید می‌کند که معادل ۳۰ درصد از کل تولید برنج ایران را تشکیل می‌دهد (Allahyari *et al*., 2016; Moghaddam & Kholghi, 2025). با توجه به وابستگی شدید این بخش

کشاورزی به منابع آب زیرزمینی و سطحی، پایش دقیق و مدل‌سازی کیفیت آب‌های زیرزمینی در این استان از اولویت بالایی برخوردار است.

جدول (۱) نشان‌دهنده گزارشی از متغیرهای کیفی آب‌های زیرزمینی استان گیلان به‌صورت روزانه برای دوره زمانی ۲۰۰۲ تا ۲۰۱۸ می‌باشد. آمار توصیفی متغیرهای کیفی این جدول شامل حداقل، حداکثر، میانگین و انحراف معیار داده‌های تمامی چاه‌های موجود است. در این مرحله از پیش‌پردازش و حذف خودکار آن‌ها برای گزارش زیر خودداری شد تا اطلاعات واقعی مرتبط با شرایط آبخوان حفظ شود.

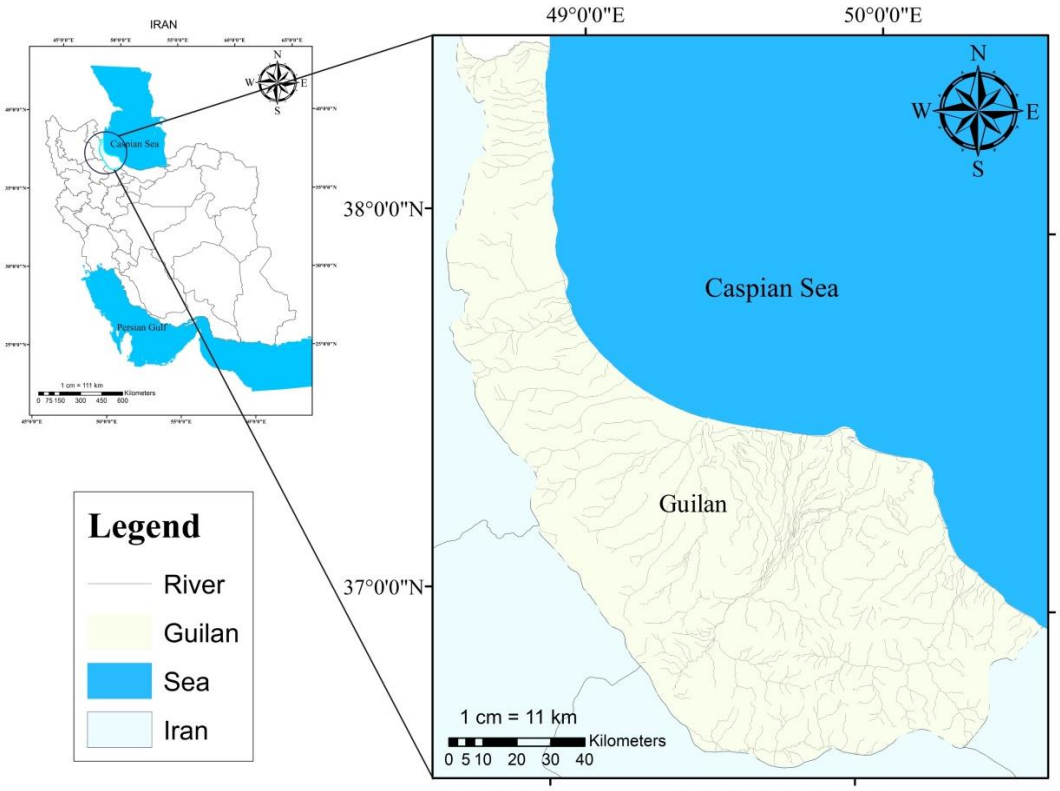


Figure 2. Location of the study area: Guilan Province, Iran.

Table 1. Statistical Summary of Groundwater Quality Parameters in Guilan Province (2002–2018)

	K^{+}	Na^{+}	Mg^{+2}	Ca^{+2}	$So4^{-2}$	Cl^{-1}	$Hco3^{-1}$	$Co3^{-2}$	PH	TDS	EC
Units	**meq/l**	**meq/l**	**meq/l**	**meq/l**	**meq/l**	**meq/l**	**meq/l**	**meq/l**		**Mg/l**	**µS/cm**
Maximum	2	111.7	17.9	30	41.6	117.6	25.1	0.7	8.8	9.3	14800
Minimum	0.01	0.02	0.2	0.6	0.03	0.1	0.9	0.03	5.2	0.1	200
Average	0.08	2.6	2.05	4.3	1.3	2.5	5.1	0.2	7.6	0.6	900
Standard deviation	0.2	6.6	1.4	2.6	2.4	6.9	2.5	0.1	0.5	0.6	900

۳. یافته‌های پژوهش

در این بخش عملکرد مدل‌ها در سناریوهای مختلف تعداد چاه‌های همسایه بررسی شد تا میزان اثرگذاری اطلاعات فضایی بر پیش‌بینی کیفیت آب زیرزمینی مشخص شود. نتایج نشان می‌دهد که افزایش تعداد چاه‌های مجاور لزوماً باعث بهبود یکسان عملکرد همه مدل‌ها نمی‌شود و این موضوع تا حد زیادی به نوع متغیر هدف وابسته است. بر همین اساس نتایج کمّی برای متغیر EC در جدول (۲) و متغیر Ca^{+2} در جدول (۳) ارائه شده است.

Table 2. Comparison of the average performance of models with 4, 5, 6, 7 and 8 neighboring wells in Guilan Province for EC

Model	R-squared (Testing)	RMSE (µS/cm) (Testing)	MAE (µS/cm) (Testing)	Pearson correlation (Testing)	Number of neighbouring Wells	Average distance from the target well (meter)
MLP	0.844	159.408	114.646	0.919	4	7053.20
KNN	0.815	173.322	125.683	0.905		
RF	0.845	158.724	113.832	0.920		
GBR	0.846	158.554	113.676	0.920		
GPR	0.846	158.580	113.710	0.920		
XGBoost	0.846	158.580	113.711	0.920		
MLP	0.856	157.821	112.186	0.926	5	7629.48
KNN	0.824	174.182	125.416	0.913		
RF	0.856	157.943	111.992	0.926		
GBR	0.857	157.493	111.831	0.927		
GPR	0.857	157.505	111.861	0.927		
XGBoost	0.857	157.505	111.861	0.927		
MLP	0.832	165.325	117.926	0.917	6	8298.65
KNN	0.816	171.928	123.539	0.911		
RF	0.851	154.858	111.500	0.925		
GBR	0.852	154.419	111.190	0.925		
GPR	0.852	154.398	111.156	0.925		
XGBoost	0.852	154.399	111.156	0.925		
MLP	0.882	123.112	86.395	0.942	7	8405.99
KNN	0.845	142.325	104.803	0.927		
RF	0.882	123.522	86.258	0.942		
GBR	0.882	123.565	86.356	0.942		
GPR	0.882	123.567	86.357	0.942		
XGBoost	0.882	123.567	86.358	0.942		
MLP	0.888	117.887	86.354	0.944	8	9494.83
KNN	0.853	134.141	99.524	0.935		
RF	0.888	117.165	84.173	0.944		
GBR	0.887	117.413	84.360	0.944		
GPR	0.887	117.413	84.359	0.944		
XGBoost	0.887	117.413	84.360	0.944		

Table 3. Comparison of the average performance of models with 4, 5, 6, 7 and 8 neighboring wells in Guilan Province for Ca^{+2}

Model	R-squared (Testing)	RMSE (meq/l) (Testing)	MAE (meq/l) (Testing)	Pearson correlation (Testing)	Number of neighbourhood Wells	Average distance from the target well (meter)
MLP	0.807	1.166	0.830	0.903	4	5628.76
KNN	0.711	1.458	0.944	0.851		
RF	0.788	1.233	0.849	0.894		
GBR	0.802	1.185	0.852	0.900		
GPR	0.799	1.196	0.831	0.897		
XGBoost	0.811	1.154	0.820	0.905		
MLP	0.635	1.515	0.954	0.850	5	6394.88
KNN	0.633	1.566	0.977	0.810		
RF	0.787	1.194	0.869	0.892		
GBR	0.778	1.216	0.871	0.889		
GPR	0.734	1.327	0.893	0.865		
XGBoost	0.790	1.183	0.862	0.894		
MLP	0.728	1.242	0.909	0.855	6	6867.54
KNN	0.482	1.797	1.100	0.757		
RF	0.689	1.374	0.938	0.840		
GBR	0.720	1.318	0.919	0.852		
GPR	0.713	1.344	0.923	0.848		
XGBoost	0.729	1.278	0.910	0.857		
MLP	0.762	1.266	0.931	0.877	7	7399.93
KNN	0.521	1.776	1.088	0.785		
RF	0.717	1.381	0.964	0.851		
GBR	0.749	1.323	0.939	0.869		
GPR	0.698	1.495	0.965	0.838		
XGBoost	0.761	1.264	0.926	0.875		
MLP	0.747	1.314	0.973	0.871	8	7934.65
KNN	0.440	1.957	1.155	0.759		
RF	0.678	1.460	1.027	0.843		
GBR	0.749	1.309	0.970	0.871		
GPR	0.745	1.321	0.978	0.870		
XGBoost	0.746	1.315	0.975	0.870		

۳. ۱. مقایسه عملکرد الگوریتم‌ها

در این مطالعه، الگوریتم‌های موردبررسی در سه گروه اصلی طبقه‌بندی شدند؛ گروه اول شامل الگوریتم‌های مبتنی بر

یادگیری جمعی شامل جنگل تصادفی (RF)، گرادیان بوستینگ (GBR) و XGBoost؛ گروه دوم شامل فرایند گاوسی (GPR) و گروه سوم شامل شبکه عصبی چندلایه (MLP) و الگوریتم همسایگان نزدیک (KNN). نتایج مربوط به متغیر شوری نشان داد که با افزایش تعداد چاه‌های همسایه، دقت پیش‌بینی مدل‌ها به‌طور قابل‌توجهی بهبود می‌یابد و بهترین عملکرد در سناریوی استفاده از هشت چاه همسایه حاصل شده است. در این حالت، مدل جنگل تصادفی با ضریب تعیین ۰/۸۸۸ مقدار RMSE برابر با ۱۱۷/۱۶ و MAE برابر با ۸۴/۱۷ دقیق‌ترین پیش‌بینی را ارائه داد. مدل‌های MLP، GBR، GPR و XGBoost نیز عملکرد بسیار نزدیکی با این مدل داشتند و مقادیر R^2 آن‌ها در حدود ۰/۸۸۷ تا ۰/۸۸۸ قرار گرفت که نشان‌دهنده توانایی بالای این الگوریتم‌ها در مدل‌سازی روابط غیرخطی و پیچیده توزیع فضایی شوری در آبخوان موردمطالعه است.

در مقابل، الگوریتم KNN در تمامی سناریوهای همسایگی عملکرد ضعیفی از خود نشان داد که بیانگر محدودیت این روش در مدل‌سازی الگوهای پیچیده فضایی-زمانی کیفیت آب زیرزمینی است. در مورد متغیر کلسیم، نتایج روند متفاوتی را نشان داد. بیش‌ترین دقت پیش‌بینی در سناریوی استفاده از چهار چاه همسایه حاصل شد و با افزایش تعداد چاه‌های همسایه، عملکرد مدل‌ها به‌تدریج کاهش یافت. در این سناریو، الگوریتم XGBoost بهترین عملکرد را با ضریب تعیین ۰/۸۱۱، مقدار RMSE برابر با ۱/۱۵۴ و MAE برابر با ۰/۸۲۰ ارائه داد. پس از آن مدل‌های MLP و GBR با عملکردی نزدیک در رتبه‌های بعدی قرار گرفتند. کاهش دقت مدل‌ها در همسایگی‌های بزرگ‌تر احتمالاً ناشی از ورود نویز فضایی و کاهش همبستگی مکانی بین چاه‌های دورتر است که می‌تواند فرایند یادگیری مدل را با چالش مواجه کند.

به‌طور کلی نتایج این پژوهش نشان می‌دهد که الگوریتم‌های مبتنی بر یادگیری جمعی، به‌ویژه RF و XGBoost، توانایی بالایی در پیش‌بینی پارامترهای کیفی آب زیرزمینی دارند. علاوه بر این، انتخاب تعداد مناسب چاه‌های همسایه نقش مهمی در بهبود عملکرد مدل‌ها ایفا می‌کند، به‌طوری‌که برای پارامتر شوری وابستگی فضایی گسترده‌تری مشاهده شد، درحالی‌که برای پارامتر کلسیم روابط فضایی در مقیاس محدودتری بهتر قابل مدل‌سازی بوده است.

Table 4. Summary of the average test performance of the models for Ca^{2+}

Wells	R-squared	RMSE (meq/l)	MAE (meq/l)	Pearson correlation
MLP	0.736	1.301	0.919	0.871
KNN	0.557	1.711	1.053	0.793
RF	0.732	1.329	0.929	0.864
GBR	0.760	1.270	0.910	0.876
GPR	0.738	1.336	0.918	0.864

Table 5. Summary of the average test performance of the models for EC

Wells	R-squared	RMSE (µS/cm)	MAE (µS/cm)	Pearson correlation
MLP	0.860	144.711	103.501	0.930
KNN	0.831	159.180	115.793	0.918
RF	0.858	142.442	101.551	0.931
GBR	0.864	142.289	101.482	0.932
GPR	0.865	142.293	101.489	0.932

Table 6. Summary of the average test performance of the models for both quality parameters

Wells	R-squared	RMSE	MAE	Pearson correlation
MLP	0.798	92.090	71.150	0.895
KNN	0.641	80.445	58.423	0.855
RF	0.795	71.885	51.240	0.898
GBR	0.812	71.779	51.196	0.904
GPR	0.801	71.814	51.203	0.898

۳. ۲. تحلیل مدل‌های برتر برای هر چاه

به‌منظور بررسی دقیق‌تر رفتار مدل‌های یادگیری ماشین در مقیاس مکانی، عملکرد الگوریتم‌های مختلف در سناریوهای همسایگی شامل چهار تا هشت حلقه چاه مجاور موردتحلیل قرار گرفت. این تحلیل با هدف ارزیابی میزان تأثیر اطلاعات فضایی چاه‌های اطراف بر دقت پیش‌بینی و هم‌چنین شناسایی الگوی عملکرد مدل‌ها در شرایط مختلف همسایگی انجام شد. نتایج حاصل از مقایسه شاخص‌های آماری نشان می‌دهد که الگوی عملکرد الگوریتم‌ها در اکثر سناریوها نسبتاً پایدار بوده و مدل‌های مبتنی بر روش‌های درختی و تقویتی عموماً دقت بالاتری نسبت به سایر روش‌ها ارائه داده‌اند.

در سناریوی چهار حلقه چاه همسایگی، عملکرد مدل‌ها نشان می‌دهد که الگوریتم‌های مبتنی بر یادگیری تقویتی و درختی از جمله GBR، RF و XGBoost توانسته‌اند الگوی تغییرات پارامترهای کیفی آب را با دقت مناسبی بازسازی کنند. در این حالت، مقادیر ضریب تعیین و ضریب همبستگی در سطح قابل‌قبولی قرار داشته و خطای پیش‌بینی نیز در مقایسه با سایر روش‌ها کم‌تر بوده است. در مقابل، الگوریتم KNN در این سناریو نیز عملکرد ضعیف‌تری نشان داد که بیانگر محدودیت این روش در مدل‌سازی روابط پیچیده و غیرخطی میان چاه‌های همجوار است.

با افزایش تعداد چاه‌های همسایه به پنج حلقه، روند عملکرد مدل‌ها تا حد زیادی مشابه سناریوی قبل باقی می‌ماند. در این حالت نیز مدل‌های درختی و تقویتی همچنان در میان دقیق‌ترین روش‌ها قرار دارند و اختلاف عملکرد میان آن‌ها نسبتاً محدود است. در مقابل، مدل‌های مبتنی بر فاصله مانند KNN همچنان کم‌ترین دقت را نشان می‌دهند. شبکه عصبی MLP در این سناریو عملکردی متوسط ارائه داده و اگرچه نسبت به KNN دقیق‌تر است، اما در بسیاری از موارد به سطح دقت مدل‌های تقویتی نمی‌رسد.

در سناریوی شش حلقه چاه همسایگی، رقابت میان مدل‌های درختی و تقویتی نزدیک‌تر می‌شود و اختلاف عملکرد آن‌ها در معیارهای ارزیابی بسیار محدود است. در این شرایط، مدل‌های RF، GBR و XGBoost تقریباً نتایجی مشابه ارائه داده و قادر بوده‌اند تغییرات مکانی پارامترهای کیفی آب را با دقت قابل‌قبولی پیش‌بینی کنند. مدل GPR نیز در برخی موارد عملکردی نزدیک به این گروه نشان می‌دهد که بیانگر توانایی این روش در مدل‌سازی روابط مکانی پیوسته است.

در سناریوهای هفت و هشت حلقه چاه همسایگی، دقت پیش‌بینی مدل‌ها برای برخی پارامترها افزایش یافته و رقابت میان الگوریتم‌های برتر بسیار فشرده می‌شود. در این شرایط، مدل‌های مبتنی بر یادگیری جمعی مانند RF، GBR و XGBoost اغلب دقیق‌ترین نتایج را ارائه می‌دهند و اختلاف میان شاخص‌های ارزیابی آن‌ها بسیار ناچیز است. این موضوع نشان می‌دهد که این الگوریتم‌ها توانایی بالایی در بهره‌برداری از اطلاعات فضایی چاه‌های مجاور دارند و می‌توانند ساختار پیچیده و غیرخطی داده‌های هیدروژئوشیمیایی را به خوبی مدل‌سازی کنند.

در مقابل، الگوریتم KNN تقریباً در تمامی سناریوهای همسایگی پایین‌ترین مقادیر ضریب تعیین و بیش‌ترین خطای پیش‌بینی را ثبت کرده است. حساسیت بالای این روش به نویز داده‌ها و وابستگی آن به فاصله اقلیدسی ساده باعث می‌شود که در داده‌های پیچیده و چندبعدی کیفیت آب زیرزمینی عملکرد مناسبی نداشته باشد. مدل MLP نیز اگرچه در برخی سناریوها نتایج قابل‌قبولی ارائه می‌دهد، اما در مقایسه با مدل‌های مبتنی بر درخت و روش‌های تقویتی از پایداری و دقت کم‌تری برخوردار است.

به‌طورکلی، بررسی عملکرد مدل‌ها در سناریوهای مختلف همسایگی نشان می‌دهد که الگوریتم‌های مبتنی بر یادگیری جمعی و درختی از جمله RF، GBR و XGBoost بیش‌ترین سازگاری را با ساختار فضایی داده‌های کیفیت آب زیرزمینی دارند و در اکثر شرایط توانسته‌اند دقیق‌ترین پیش‌بینی‌ها را ارائه دهند. در مقابل، روش‌های مبتنی بر فاصله مانند KNN برای چنین داده‌هایی کارایی محدودتری داشته و در اغلب سناریوها دقت پایین‌تری نشان می‌دهند.

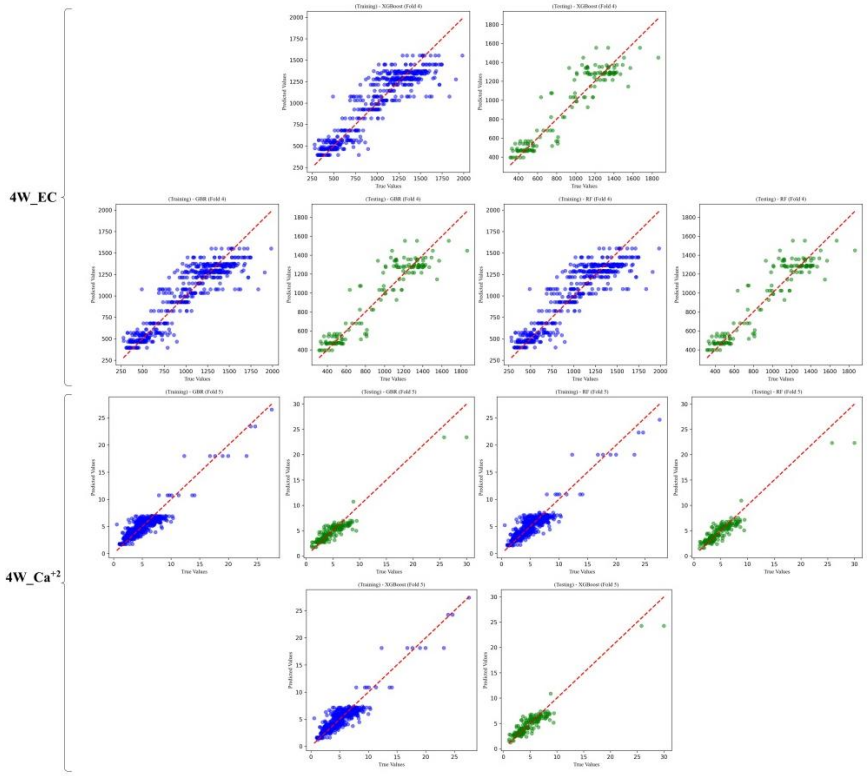


Figure 3. Best models for 4 Neighboring Wells

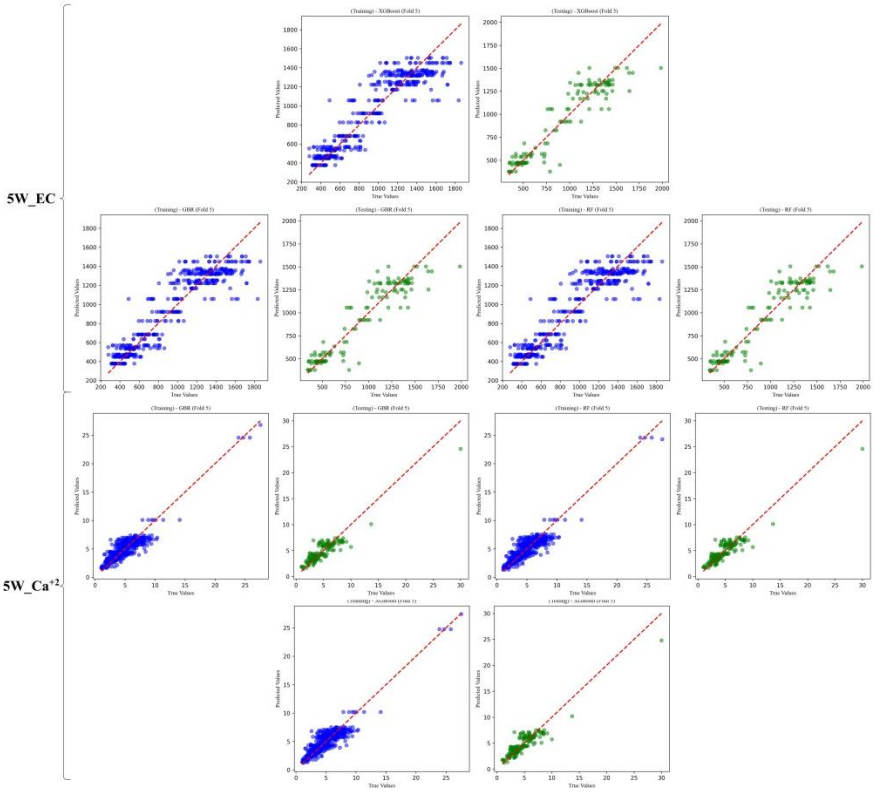


Figure 4. Best models for 5 Neighboring Wells

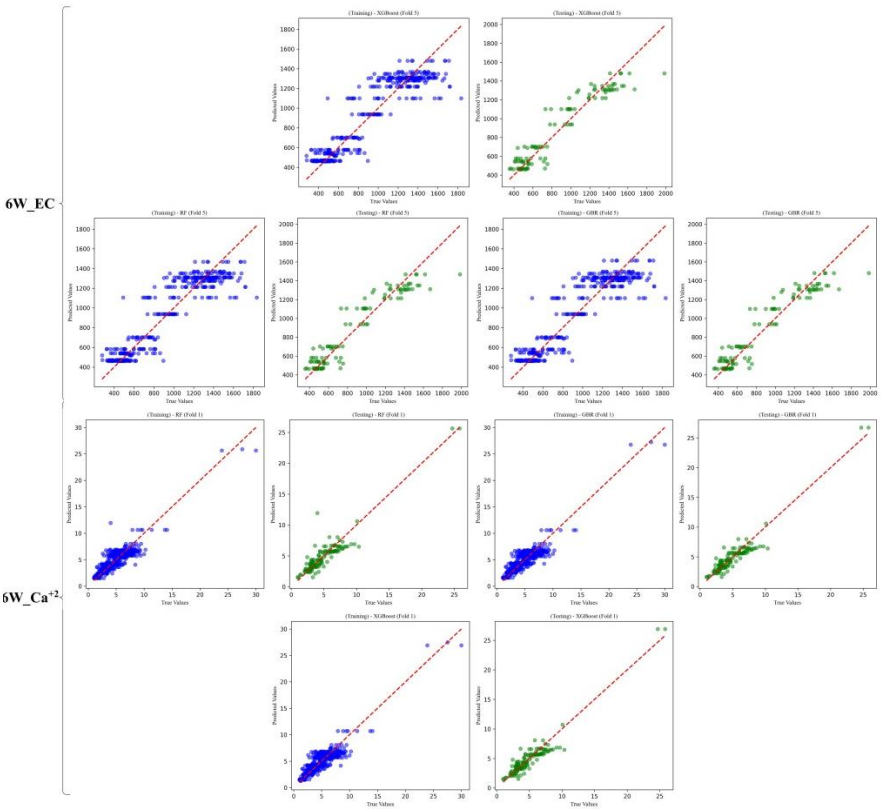


Figure 5. Best models for 6 Neighboring Wells

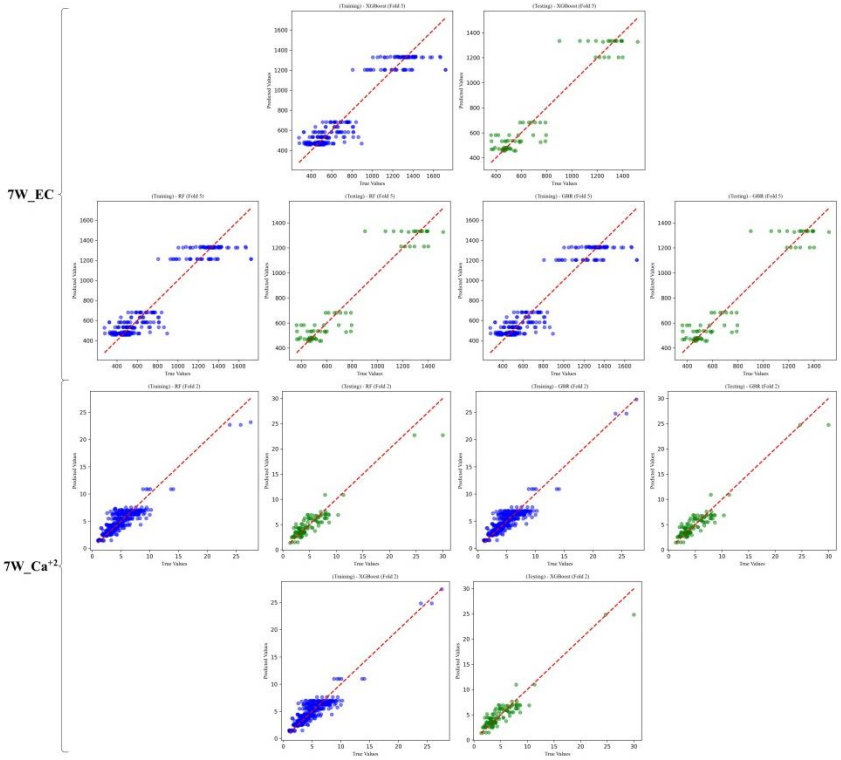


Figure 6. Best models for 7 Neighboring Wells

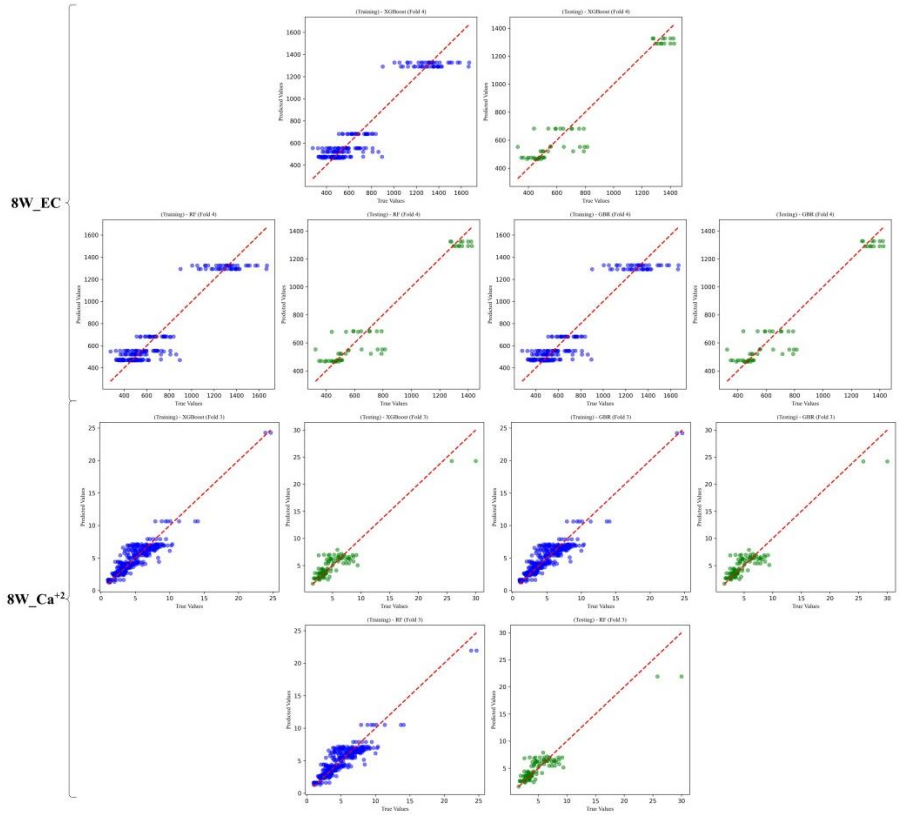


Figure 7. Best models for 8 Neighboring Wells

بررسی عملکرد مدل‌های یادگیری ماشین در پیکربندی‌های مختلف فضایی شامل چهار تا هشت چاه همجوار نشان می‌دهد که ساختار مکانی داده‌ها نقش بسیار مهمی در دقت پیش‌بینی پارامترهای کیفی آب زیرزمینی دارد. نتایج این مطالعه نشان می‌دهد که تأثیر تعداد چاه‌های همسایه بر عملکرد مدل‌ها برای پارامترهای مختلف یکسان نیست و هر پارامتر الگوی وابستگی مکانی خاص خود را نشان می‌دهد. در مورد پارامتر شوری، افزایش تدریجی تعداد چاه‌های همسایه منجر به بهبود عملکرد مدل‌ها شد و بیش‌ترین دقت در سناریوی استفاده از هشت چاه همجوار مشاهده گردید.

این روند نشان می‌دهد که تغییرات شوری در آبخوان موردمطالعه دارای همبستگی مکانی گسترده‌تری است و اطلاعات حاصل از چاه‌های دورتر نیز می‌تواند در بهبود فرایند پیش‌بینی مؤثر باشد. چنین الگویی معمولاً در سامانه‌های آبخوانی مشاهده می‌شود که در آن‌ها فرایندهایی مانند اختلاط آب‌های زیرزمینی، جریان منطقه‌ای و انتقال املاح در مقیاس بزرگ‌تر رخ می‌دهد و در نتیجه الگوهای شوری در محدوده وسیع‌تری از آبخوان قابل مشاهده است. در این شرایط، افزایش تعداد چاه‌های همسایه موجب می‌شود مدل بتواند تصویر کامل‌تری از توزیع فضایی شوری در منطقه به‌دست آورد. در مقابل، برای پارامتر کلسیم رفتار متفاوتی مشاهده شد. در این مورد، بهترین عملکرد مدل‌ها در سناریوی استفاده از چهار چاه همسایه به‌دست آمد و با افزایش تعداد چاه‌های همجوار، دقت پیش‌بینی به‌تدریج کاهش یافت. این الگو نشان می‌دهد که تغییرات کلسیم در مقیاس محلی‌تری کنترل می‌شود و همبستگی مکانی آن نسبت به شوری محدودتر است.

از دیدگاه هیدروژئوشیمیایی، غلظت کلسیم در آب‌های زیرزمینی معمولاً بیش‌تر تحت تأثیر فرایندهای محلی نظیر انحلال کانی‌های کربناته، ویژگی‌های لیتولوژیکی و شرایط ژئوشیمیایی موضعی قرار دارد (Deng *et al.*, 2020). در

نتیجه، داده‌های حاصل از چاه‌های نزدیک‌تر اطلاعات مرتبط‌تری برای مدل فراهم می‌کنند، درحالی‌که اضافه‌شدن داده‌های چاه‌های دورتر می‌تواند موجب ورود نویز فضایی و کاهش دقت پیش‌بینی شود. در مجموع، نتایج این تحلیل نشان می‌دهد که تعیین تعداد بهینه چاه‌های همسایه برای مدل‌سازی کیفیت آب زیرزمینی به نوع پارامتر موردبررسی وابسته است. درحالی‌که پارامترهایی با الگوی انتشار منطقه‌ای مانند شوری از استفاده از همسایگان بیش‌تر بهره می‌برند، پارامترهایی که تحت تأثیر فرایندهای ژئوشیمیایی محلی قرار دارند، با تعداد محدودتری از چاه‌های مجاور بهتر قابل پیش‌بینی هستند. این یافته اهمیت درنظرگرفتن ویژگی‌های فضایی و ماهیت هیدروژئوشیمیایی پارامترها را در طراحی مدل‌های پیش‌بینی کیفیت آب زیرزمینی برجسته می‌کند.

یافته‌های این پژوهش مبنی بر تأثیر مثبت لحاظ‌کردن ساختار مکانی و چاه‌های همجوار بر دقت مدل‌ها، با نتایج مطالعات پیشین همچون مقدم و همکاران (۲۰۲۵) و عاقلی و همکاران (۲۰۲۴) که بر اهمیت حیاتی ورودی‌های مبتنی بر فاصله در بهبود عملکرد الگوریتم‌هایی مانند جنگل تصادفی تأکید داشتند، کاملاً همسو است. با این‌حال، نتیجه کلیدی این مطالعه مبنی بر «تفاوت مقیاس بهینه همسایگی برای پارامترهای مختلف» (هشت چاه برای EC در مقابل چهار چاه برای کلسیم)، رویکردی فراتر از پژوهش‌های متداول ارائه می‌دهد. بسیاری از مطالعات گذشته (مانند عیسی‌زاده و همکاران، ۲۰۱۷؛ کومار و همکاران، ۲۰۲۰) وابستگی مکانی را به‌صورت یکنواخت برای همه پارامترهای کیفی در نظر گرفته یا تنها بر یک شاخص متمرکز شده بودند. یافته حاضر با تأیید دیدگاه دنگ و همکاران (۲۰۲۰) در خصوص کنترل محلی فرایندهای ژئوشیمیایی کلسیم، نشان می‌دهد که تعمیم یک الگوی مکانی ثابت و یکسان برای تمامی پارامترها (آن‌طور که در برخی مدل‌سازی‌های پیشین فرض شده است) می‌تواند منجر به ورود نویز فضایی و کاهش دقت پیش‌بینی شود. بدین ترتیب، این پژوهش با اثبات وابستگی مقیاس بهینه به ماهیت هیدروژئوشیمیایی هر پارامتر، شکاف موجود در درک پیچیدگی‌های وابستگی مکانی را پر کرده و با نتایج ساده‌انگارانه پیشین که تفاوت مقیاس‌های مکانی را نادیده می‌گرفتند، تمایز دارد.

Table 7. Summary of the average test performance for each number of neighboring wells for Ca^{2+}

Wells	R-squared	RMSE (meq/l)	MAE (meq/l)	Pearson correlation
4	0.813	81.398	58.491	0.903
5	0.742	81.111	57.743	0.897
6	0.760	80.307	57.680	0.878
7	0.789	64.014	45.195	0.894
8	0.783	60.842	44.101	0.895

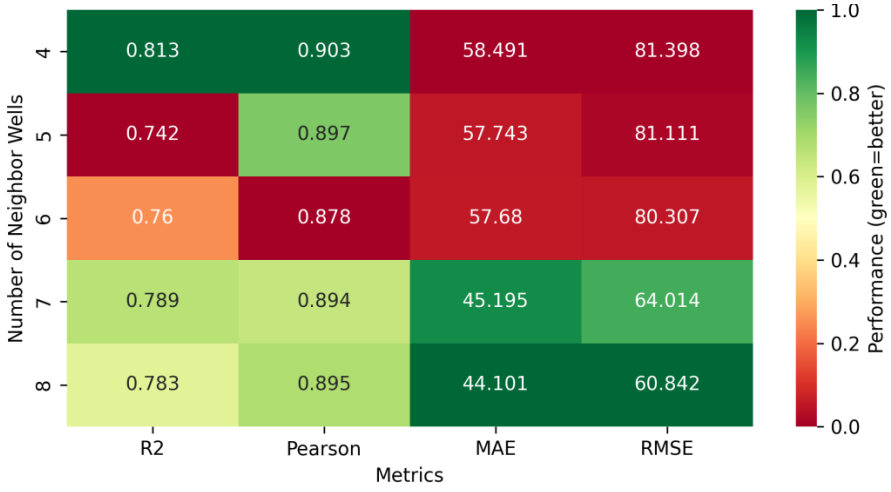


Figure 8. Heatmap of the mean test performance for each number of neighboring wells for Ca^{2+}

Table 8. Summary of the mean test performance for each number of neighboring wells for EC

Wells	R-squared	RMSE (µS/cm)	MAE (µS/cm)	Pearson correlation
4	0.840	161.194	115.876	0.918
5	0.851	160.408	114.191	0.924
6	0.842	159.221	114.411	0.921
7	0.875	126.610	89.421	0.939
8	0.882	120.239	87.189	0.943

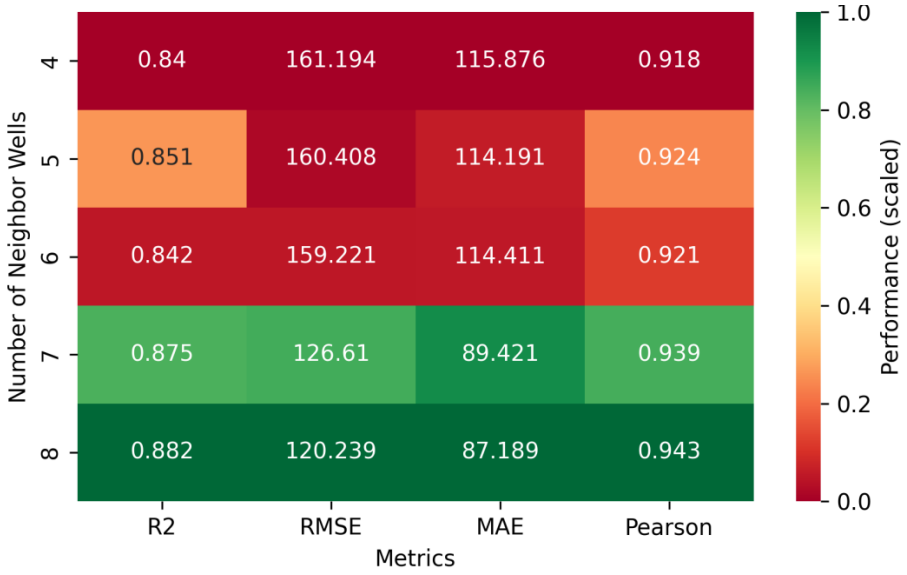


Figure 9. Heatmap of the average test performance for each number of neighboring wells for EC

Table 9. Summary of the mean test performance for each number of neighboring wells for both variables

Wells	R-squared	RMSE	MAE	Pearson correlation
4	92.090	71.150	0.895	0.798
5	80.445	58.423	0.855	0.641
6	71.885	51.240	0.898	0.795
7	71.779	51.196	0.904	0.812
8	71.814	51.203	0.898	0.801

۴. بحث و نتیجه‌گیری

در این پژوهش، کارایی مجموعه‌ای از الگوریتم‌های پیشرفته یادگیری ماشین شامل Random Forest (RF)، XGBoost، Gradient Boosting Regression (GBR)، Gaussian Process Regression (GPR)، شبکه عصبی چندلایه (MLP) و K-Nearest Neighbors (KNN) برای مدل‌سازی و پیش‌بینی پارامترهای کیفیت آب زیرزمینی موردبررسی قرار گرفت. به‌منظور افزایش دقت مدل‌ها، رویکرد مهندسی ویژگی‌های فضایی به‌کار گرفته شد که در آن اطلاعات مکانی چاه‌های مجاور به‌عنوان متغیرهای ورودی به مدل اضافه شد. در این چارچوب، پنج سناریوی مختلف فضایی شامل چهار تا هشت چاه همسایه مورد ارزیابی قرار گرفت تا اثر ساختار فضایی داده‌ها بر عملکرد الگوریتم‌ها مشخص شود.

نتایج حاصل از تحلیل مدل‌ها نشان داد که الگوریتم‌های مبتنی بر یادگیری جمعی[۳۲] عملکرد به مراتب بهتری نسبت به سایر روش‌ها دارند. در میان الگوریتم‌های موردبررسی، مدل‌های Random Forest، XGBoost و GBR توانستند روابط غیرخطی و پیچیده میان متغیرهای هیدروشیمیایی و ویژگی‌های فضایی را با دقت بیش‌تری مدل‌سازی کنند. در مقابل، الگوریتم KNN به‌دلیل اتکای صرف به فاصله‌های ساده در فضای ویژگی‌ها و مدل MLP به‌دلیل حساسیت بالا به تنظیم پارامترها، عملکرد ضعیف‌تری از خود نشان دادند.

تحلیل جداگانه پارامترهای هیدروشیمیایی نشان داد که الگوی وابستگی فضایی پارامترها یکسان نیست. برای شوری آب زیرزمینی، افزایش تعداد چاه‌های همسایه موجب بهبود قابل‌توجه عملکرد مدل‌ها شد و بهترین نتیجه در سناریوی هشت چاه همسایه حاصل گردید. در این حالت، مدل Random Forest با مقدار R^2 برابر با ۰/۸۸۸ کم‌ترین خطای پیش‌بینی، بالاترین توانایی را در تبیین تغییرات شوری نشان داد. این موضوع بیانگر آن است که تغییرات شوری در منطقه موردمطالعه تحت تأثیر فرایندهای هیدروژئوشیمیایی در مقیاس منطقه‌ای قرار دارد و اطلاعات مکانی گسترده‌تر می‌تواند دقت پیش‌بینی را افزایش دهد.

در مقابل، برای پارامتر کلسیم (Ca^{+2}) بهترین عملکرد در سناریوی چهار چاه همسایه مشاهده شد که در آن مدل XGBoost با مقدار R^2 برابر با ۰/۸۱۱ بالاترین دقت پیش‌بینی را نشان داد. این نتیجه نشان می‌دهد که توزیع مکانی کلسیم بیش‌تر تحت تأثیر فرایندهای ژئوشیمیایی محلی مانند انحلال کانی‌ها و برهم‌کنش آب و سنگ قرار دارد و وابستگی فضایی آن در مقیاس محدودتری شکل می‌گیرد. با این‌حال، زمانی که عملکرد کلی مدل‌ها برای هر دو پارامتر به‌صورت هم‌زمان موردبررسی قرار گرفت، نتایج هیت‌مپ ترکیبی نشان داد که سناریوی استفاده از هفت چاه همسایه بهترین تعادل میان دقت مدل و میزان خطا را فراهم می‌کند. در این سناریو، مقدار همبستگی پیرسون به بیش‌ترین سطح خود رسید و مقادیر خطا نیز در پایین‌ترین محدوده پایدار قرار داشتند. بنابراین می‌توان نتیجه گرفت که هفت چاه همسایه به‌عنوان سناریوی بهینه فضایی برای مدل‌سازی هم‌زمان پارامترهای کیفیت آب زیرزمینی در منطقه موردمطالعه قابل پیشنهاد است.

در مجموع، نتایج این پژوهش نشان داد که RF در سناریوی هشت چاه همسایه برای EC و XGBoost در سناریوی چهار چاه همسایه برای Ca بهترین عملکرد را داشته‌اند. این تفاوت بیانگر آن است که تعداد بهینه چاه‌های همسایه باید با توجه به نوع متغیر کیفی تعیین شود و نمی‌توان یک الگوی ثابت را برای همه پارامترها به‌کار گرفت. افزون بر این، نتایج به‌دست‌آمده نشان می‌دهد که ترکیب الگوریتم‌های یادگیری ماشین مبتنی بر مدل‌های جمعی با مهندسی ویژگی‌های فضایی، می‌تواند چارچوبی کارآمد برای پیش‌بینی کیفیت آب‌های زیرزمینی فراهم کند. این رویکرد علاوه بر آنکه توانایی بازنمایی ساختارهای پیچیده فضایی–زمانی پارامترهای هیدروشیمیایی را دارد، می‌تواند در پایش تغییرات کیفی آبخوان‌ها، شناسایی نواحی در معرض خطر و پشتیبانی از تصمیم‌گیری‌های مدیریتی در راستای بهره‌برداری پایدار از منابع آب زیرزمینی نیز مورداستفاده قرار گیرد.

در نهایت، باید تأکید شود که مقادیر عددی بهینه (مانند چهار یا هشت چاه همسایه) به‌هیچ‌وجه مطلق نبوده و تعمیم‌پذیری مستقیم به سایر آبخوان‌ها را ندارند، چراکه این اعداد به‌شدت به ویژگی‌های بومی منطقه (نظیر چگالی شبکه پایش و ناهمگنی هیدرولوژیکی آبخوان) وابسته‌اند. با این‌حال، چارچوب روش‌شناختی و اصل مفهومی پژوهش مبنی بر «تفاوت مقیاس بهینه براساس ماهیت هیدروژئوشیمیایی پارامترها» کاملاً تعمیم‌پذیر است. بنابراین پیشنهاد می‌شود در مطالعات آتی برای سایر مناطق، تعداد بهینه همسایگان از طریق تحلیل حساسیت و با در نظر گرفتن شرایط بومی همان آبخوان به‌صورت موردی تعیین گردد.

۵. محدودیت‌ها و پژوهش‌های آتی

اگرچه در پژوهش حاضر تمرکز اصلی بر ارزیابی مستقل اثر ساختار فضایی و مهندسی ویژگی‌های مبتنی بر چاه‌های همجوار بود، اما یکی از محدودیت‌های این مطالعه، عدم استفاده از سایر متغیرهای پیش‌بین محیطی و هیدروژئولوژیکی است. مطالعات پیشین نشان داده‌اند که تلفیق داده‌های مکانی با متغیرهای توپوگرافی (مانند مدل رقومی ارتفاع، شیب و فاصله از آبراهه‌ها)، کاربری اراضی و پارامترهای اقلیمی (نظیر میزان بارش و تبخیر) می‌تواند دقت مدل‌های یادگیری

ماشین را در پیش‌بینی کیفیت آب زیرزمینی به طرز چشم‌گیری افزایش دهد (Zhang *et al*., 2021; Chen *et al*., 2021). هم‌چنین، استفاده از سایر پارامترهای فیزیکوشیمیایی به‌عنوان متغیرهای ورودی در بسیاری از پژوهش‌ها مرسوم است. با این‌حال، در این مطالعه بیش‌تر به این دلیل از ورود این متغیرها اجتناب شد که هدف اصلی، سنجش خالص «همبستگی مکانی» بدون دخالت متغیرهای مخدوش‌کننده محیطی بود. افزون بر این، در بسیاری از شبکه‌های پایش واقعی، به‌ویژه در مناطق با محدودیت داده، دسترسی به داده‌های دقیق توپوگرافی یا آنالیز کامل هیدروشیمیایی دشوار است. از این‌رو، مدل‌های مبتنی بر داده‌های مکانی چاه‌های مجاور، رویکردی کاربردی و کم‌هزینه برای مدیریت منابع آب ارائه می‌دهند. بنابراین، پیشنهاد می‌شود در پژوهش‌های آتی، با تلفیق ویژگی‌های فضاییِ به‌دست‌آمده در این مطالعه با متغیرهای محیطی و اقلیمی مؤثر در منطقه (مانند مطالعات Dinpashoh *et al*. (2019) و Behzadi (2021) در استان گیلان)، مدل‌های جامع‌تری برای پیش‌بینی کیفیت آب زیرزمینی توسعه داده شود.

۶. پی‌نوشت‌ها

1. Hydro chemical Modeling
2. IWRM
3. Artificial Neural Network
4. Machine Learning
5. Dou *et al*.
6. Hulu Watershed
7. Kumar *et al*.
8. Artificial Neural Network
9. Zhang *et al*.
10. Random Forest
11. Outliers
12. Liu *et al*.
13. Gradient Boosting Regression
14. Extreme Gradient Boosting
15. Chen *et al*.
16. Fuhe
17. Multilayer Perceptron
18. Rectified Linear Unit
19. Overfitting
20. K-Nearest Neighbors
21. Ensemble
22. Estimators
23. Boosting
24. n_estimators
25. Objective Function
26. MSE
27. Gaussian Process Regressor
28. K-Fold Cross-Validation
29. Training
30. Testing
31. Sensitivity Analysis
32. Ensemble Learning

۷. تعارض منافع

هیچ‌گونه تعارض منافعی بین نویسندگان وجود ندارد.

۸. منابع

معاونت برنامه‌ریزی و توسعه منابع آب. (۱۴۰۰). گزارش سالانه مدیریت منابع آب کشور. وزارت نیرو.

Agheli, E., Ashrafzadeh, A., & Janatrostami, S. (2024). Modeling the spatial dependence of quality data using distance-based inputs and a data-driven model. Groundwater for Sustainable Development, 24, 101061. https://doi.org/10.1016/j.gsd.2023.101061

Alijani, B., Ghasemi, M., & Hosseini, M. (2018). Climate change and water resources in Iran: A review. Iranian Journal of Geophysics, 12(1), 1–15.

Allahyari, M. S., Ghavami, S., Daghighi Masuleh, Z., Michailidis, A., & Nastis, S. A. (2016). Understanding farmers' perceptions and adaptations to precipitation and temperature variability: Evidence from Northern Iran. Climate, 4(4), 58. https://doi.org/10.3390/cli4040058

Behzadi, J. (2021). Distribution of water shortage and soil moisture consumption (Case study: Guilan Province). Water Harvesting Research, 4(2), 238–244. https://doi.org/10.22077/jwhr.2022.5042.1053

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

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). ACM.

Chen, Y., Wang, J., & Li, X. (2021). Predicting river water quality using extreme gradient boosting model under climate change scenarios. Science of the Total Environment, 785, 147289. https://doi.org/10.1016/j.scitotenv.2021.147289

Dalmasso, M., Civera, M., De Biagi, V., *et al*. (2025). Gaussian process regression (GPR)-based missing data imputation and its uses for bridge structural health monitoring. ABEN, 6, 23. https://doi.org/10.1186/s43251-025-00169-1

Deng, Y., Deng, J., Xu, B., Luo, Y., & Zhang, Q. (2020). Hydrogeochemical characteristics and controlling factors of groundwater in a karst area of Southwest China. Journal of Hydrology, 590, 125561.

Dinpashoh, Y., Singh, V. P., Biazar, S. M., & Kavehkar, S. (2019). Impact of climate change on streamflow timing (Case study: Guilan Province). Theoretical and Applied Climatology, 138, 65–76. https://doi.org/10.1007/s00704-019-02810-2

Dou, G., Chen, R., Han, C., Liu, Z., & Liu, J. (2022). Research on water-level recognition method based on image processing and convolutional neural networks. Water, 14(12), 1890. https://doi.org/10.3390/w14121890

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Gurdak, J. J., & Qi, S. L. (2012). Updated conceptual model and conceptual numerical model for groundwater vulnerability to anthropogenic impacts, Yucca Mountain, Nevada, USA. Environmental Earth Sciences, 65(7), 1955–1975. https://doi.org/10.1007/s12665-011-1273-8

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Isazadeh, M., Biazar, S. M., & Ashrafzadeh, A. (2017). Support vector machines and feed-forward neural networks for spatial modeling of groundwater qualitative parameters. Environmental Earth Sciences, 76, 1–14. https://doi.org/10.1007/s12665-017-6938-5

Isaaks, E. H., & Srivastava, R. M. (1989). An introduction to applied geostatistics. Oxford University Press.

Kallab, L., Mansour, E., & Chbeir, R. (2024). Towards ML models' recommendations. Data Science and Engineering, 9, 409–430. https://doi.org/10.1007/s41019-024-00262-x

Kumar, P., Singh, R., & Jain, A. K. (2020). Application of artificial neural networks for groundwater quality prediction: A case study. Journal of Hydrology, 585, 124812. https://doi.org/10.1016/j.jhydrol.2020.124812

Liu, Y., Wang, Y., & Zhang, X. (2022). Predicting nitrate concentration in groundwater using extreme gradient boosting (XGBoost) model. Science of the Total Environment, 804, 150123. https://doi.org/10.1016/j.scitotenv.2021.150123

Moghaddam, N., & Tavana, A. H. (2023, November). An overview of the effects of climate change on surface and underground water sources. In Proceedings of the Second National Conference on Water Quality Management and the Fourth National Conference on Water Consumption Management. Technical Schools of Tehran University, Tehran, Iran.(In Persian).

Moghaddam, N., & Kholghi, M. (2025). Analysis of Groundwater Table Decline and Salinity Intensification in the Qazvin Plain: Implications from a Water Resources Governance Perspective. Iranian Journal of Irrigation & Drainage, 19(3), 469-487.(In Persian).

Moghaddam, N., Hadizadeh, H., & Ashrafzadeh, A. (2025). Data-driven analysis for estimating and predicting well salinity using artificial intelligence algorithms. Discover Applied Sciences, 8, 114.

Moghaddam, N., & Ashrafzadeh, A. (2026). Groundwater salinity modeling and prediction in Qazvin Plain using a data-driven and artificial intelligence approach. Iran-Water Resources Research, 21(4), 158–175(In Persian).

Singh, V. K., Pandey, R. P., & Mishra, V. N. (2019). Machine learning approaches in hydrology: A review. Water Resources Management, 33(10), 3379–3400. https://doi.org/10.1007/s11269-019-02318-7

Su, W., Jiang, F., Shi, C., *et al*. (2023). An XGBoost-based knowledge tracing model. International Journal of Computational Intelligence Systems, 16, 13. https://doi.org/10.1007/s44196-023-00192-y

Wada, Y., Wisser, D., & Bierkens, M. F. (2010). Modeling global groundwater depletion rates. Journal of Geophysical Research: Atmospheres, 115(D20). https://doi.org/10.1029/2010JD014568

Zhang, Y., Chen, N., & Li, H. (2021). Random forest-based groundwater quality assessment in a complex aquifer system. Environmental Monitoring and Assessment, 193, 1–15. https://doi.org/10.1007/s10661-021-09123-4