پیش‌بینی مقدار کل جامدات حل شده در آب زیرزمینی با استفاده از دو روش یادگیری ماشین و شبکه عصبی

مبین وکیلی
دانشگاه صنعتی امیرکبیر(پلی‌تکنیک تهران)
تقی عبادی
دانشگاه صنعتی امیرکبیر(پلی‌تکنیک تهران)
امیر گل رو
دانشگاه صنعتی امیرکبیر(پلی‌تکنیک تهران)

چکیده

در سال های اخیر آب زیرزمینی به عنوان بزرگترین منبع آب شیرین در دسترس بشر، از لحاظ کیفی و کمی دچار مشکلاتی شده است. از این رو محققین راه های متنوعی را برای پاکسازی این منبع مهم پیشنهاد داده اند که یکی از این راه ها روش الکتروکینتیک میباشد. جهت طراحی روش الکتروکینتیک باید مشخصاتی از آبخوان مانند هدایت الکتریکی مورد بررسی قرار گیرد که یکی از راه های شناخت هدایت الکتریکی، مقدار کل جامدات حل شده است. در این تحقیق با استفاده از روش های یادگیری ماشین و شبکه عصبی سعی شده تا مدلی بهینه و دقیق برای پیشبینی مقدار کل جامدات حل شده در آب زیرزمینی ساخته شود. برای ساخت مدل مورد بحث از یک پایگاه داده مربوط به آب زیرزمینی آمریکا که بین سال های ۱۹۹۱ الی ۲۰۱۸ جمع آوری شده، استفاده شده است. در نتایج این تحقق نشان داده شد که هر دو روش یادگیری ماشین و شبکه عصبی از لحاظ دقت، عملکرد مشابهی ارائه دادند و ضریب تعیین ۰.۹۳ برای داده های آزمون بدست آمد. هردو مدل از لحاظ بیش برازش و کم برازش بودن نیز مورد بررسی قرار گرفتند. در نهایت نیز دیده شد که روش یادگیری ماشین از لحاظ مصرف منابع پردازشی بهینه تر بوده و توانسته با مقدار کمتری از مصرف پردازنده به همان عملکرد شبکه عصبی برسد.

واژگان کلیدی: آب زیرزمینی، علوم داده، یادگیری ماشین، شبکه عصبی، کل جامدات حل شده، کیفیت آب

مقدمه

آب زیرزمینی به‌عنوان یکی از منابع حیاتی برای بشر نقشی اساسی در زندگی او ایفا کرده و یکی از پارامترهای تأثیرگذار در توسعه پایدار می‌باشد (Programme, 2022). اصولاً آب زیرزمینی به آبی گفته می‌شود که در ترازی پایین‌تر از سطح زمین، فضای خالی بین دانه‌های خاک را پر می‌کند. این بخش از خاک که حاوی آب زیرزمینی می‌باشد کاملاً اشباع بوده و تنها شامل آب، دانه‌های خاک و درصد کمی هوا (به دلیل غیرممکن‌بودن رسیدن به اشباع کامل در طبیعت) می‌باشد (Freeze R.A. , 1979). آب زیرزمینی به‌عنوان یکی از منابع اصلی بشر برای تأمین آب شیرین، نیازهای میلیون‌ها نفر را در دنیا از جمله آشامیدن، کشاورزی، صنعت و... را برطرف کرده و باعث پیشرفت و بقای جوامع بشری می‌شود (Nejatijahromi et al., 2019). از طرفی نیز، با افزایش جمعیت و گسترش مدرنیته، رشد نیازهای بشر سیری نمایی داشته و باعث آلودگی این منبع خدادادی شده است. اهمیت آب‌های زیرزمینی تنها به این موارد محدوده نشده و اهداف زیادی از اهداف سازمان ملل نیز بر اهمیت و تأثیر آن بر زندگی بشر و سایر موجودات کره زمین صحه می‌گذارد. شاخص‌ترین هدف از اهداف توسعه پایدار سازمان ملل، هدف شماره ۶ یا دستیابی به آب تمیز و بهداشتی (SDG 6-Clean Water and Sanitation) می‌باشد؛ زیرا آب زیرزمینی به‌عنوان یکی از منابع در دسترس و عمومی در تمام مناطق کره زمین، می‌تواند تأمین‌کننده آب موردنیاز جوامع محلی که دسترسی به شبکه توزیع آب ندارند باشد (Programme, 2022).

آب زیرزمینی مشخصات شیمیایی و فیزیکی بسیار متفاوتی دارد که هرکدام در یکی از کاربردهای محیط‌زیستی، صنعتی و... استفاده می‌شوند (Zhang et al., 2017). مشخصاتی مانند غلظت انواع آلاینده، pH و... در محیط‌زیست مهم بوده و ممکن است در صورت بالابودن مقادیر ذکر شده، روش‌هایی جهت رفع این مشکلات پیش‌بینی شود (Ghaeminia & Mokhtarani, 2018). یکی از مشخصات آب زیرزمینی که هم در صنعت و هم در محیط‌زیست حائز اهمیت است مقدار کل جامدات حل شده (TDS) می‌باشد. مواد جامد محلول کل معیاری اساسی در ارزیابی کیفیت آب‌های زیرزمینی است که تناسب آن برای اهداف مختلف، از آب آشامیدنی تا آبیاری، تأثیرگذار است (Priyanka et al., 2016). TDS که به‌عنوان مجموع تمام مواد معدنی و آلی حل‌شده در آب تعریف می‌شود، تصویری اساسی از محتوای معدنی و شوری آب ارائه می‌دهد. درک پویایی از TDS در آب‌های زیرزمینی برای تضمین شیوه‌های پایدار مدیریت آب و حفاظت از سلامت عمومی ضروری است (Amjad Aliewi, 2020). این مقدار که در آب زیرزمینی نسبت به آب سطحی بالاتر است روی استفاده آب در صنعت و فرایندهای حساس تأثیرگذار بوده و در کنار آن روی هدایت الکتریکی آب نیز مؤثر است (Zhou Xun, 2007). تحقیقات همچنان به بررسی تعاملات پیچیده بین TDS، سایر پارامترهای کیفیت آب و سلامت انسان ادامه می‌دهد. شناسایی خطرات بالقوه سلامتی مرتبط با اجزای خاص درون TDS همچنان یک مسیر تحقیقاتی در حال انجام است (Yujie Ji, 2020). علاوه بر این، توسعه فناوری‌های مقرون‌به‌صرفه تصفیه TDS و شیوه‌های پایدار مدیریت آب، اولویت‌های کلیدی برای تضمین امنیت آب در آینده هستند.

بیان مسئله:

به شکل کلی استفاده از مدل های یادگیری ماشین در تخمین تراز و سایر مشخصات هدرولوژیک آب زیر زمینی تاریخچه طولانی دارد اما در سال های اخیر روش های هوش مصنوعی متنوعی نیز برای پیش‌بینی پارامتر های متنوع آب زیرزمینی استفاده شده است (ترابی و همکاران ۱۳۹۸؛ دانشور وثوقی و منافیان ۱۳۹۷). به عنوان مثال (عزیزپور و همکاران ۱۴۰۰) چند مدل هوش مصنوعی را روی داده های یک چاه مشاهده آموزش دادند که به دلیل تعداد کم داده ها برای مدل های شبکه عصبی میتواند مشکل ساز باشد. در تمام این تحقیقات مقایسه کاملی بین روش های یادگیری ماشین و شبکه عصبی انجام نشده است و مقایسه این دو روش برای حل محققین در آینده بسیار کارگشا خواهد بود. در تحقیق حاضر با استفاده از مدل های یادگیری ماشین و شبکه عصبی مصنوعی سعی شده که مدلی برای پیش‌بینی مقدار کل جامدات حل شده در آب زیرزمینی آموزش داده شود. در نهایت نیز این مدل ها بر اساس پارامتر های آماری و مصرف منابع پردازشی مورد مقایسه قرار گرفتند.

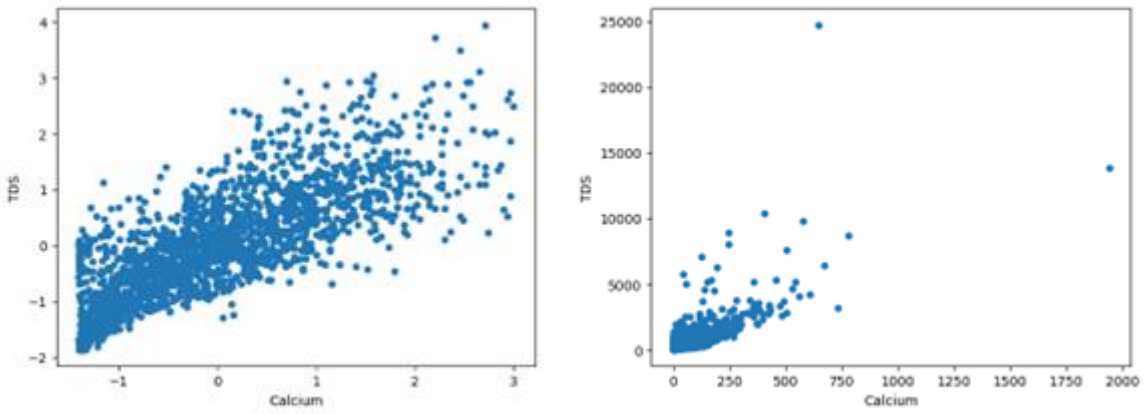
روش تحقیق:

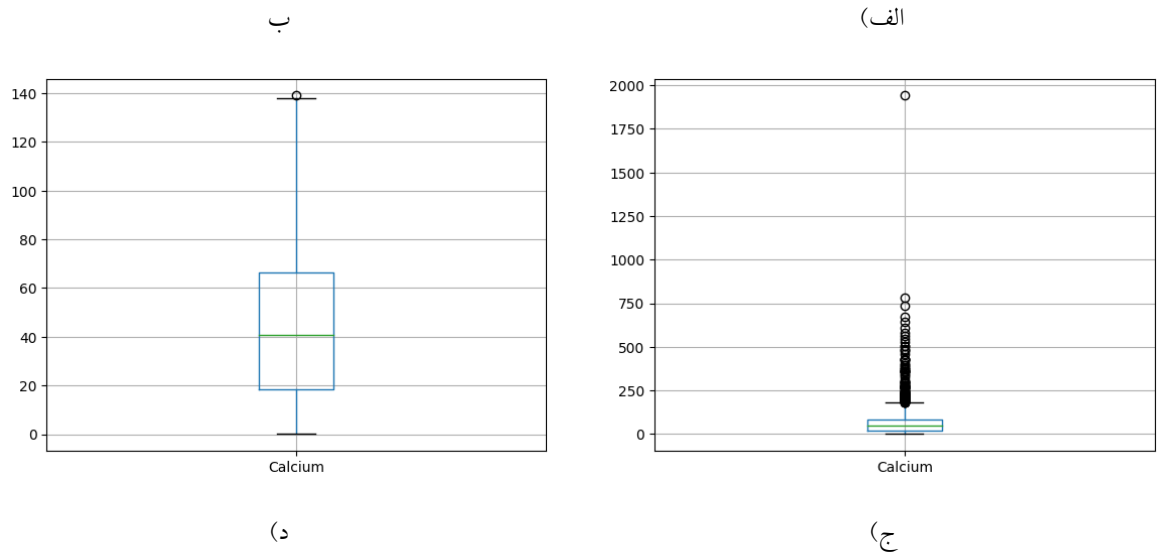
در این پژوهش داده‌های مربوط به مشخصات آب زیرزمینی در آمریکا که بین سال‌های ۱۹۹۱ الی ۲۰۱۸ در توسط سازمان زمین‌شناسی آمریکا و در سراسر پهنه جغرافیایی این کشور جمع‌آوری‌شده، استفاده شده است (Musgrove, 2020). این داده‌ها شامل ۴۸۲۴ سطر و ۵۰ ستون بوده که مشخصاتی مانندی نوع و نام آبخوان، غلظت عناصر مختلف در آب و سایر مشخصات فیزیکی، هیدرولوژیکی و شیمیایی را شامل می‌شود. از پارامترهای موجود در این دیتابیس، ۸ پارامتر کیفی و سایر کمی می‌باشند. این دیتابیس در برخی نقاط دارای مقادیر خالی یا پرت می‌باشند که در مرحله پاک‌سازی داده به شکل کامل بررسی شده‌اند. در تحقیق حاضر برای بررسی پارامترها و همچنین آموزش مدل و ارزیابی آن از زبان برنامه‌نویسی پایتون نسخه ۳.۸ استفاده شده است. تعدادی از کتابخانه‌های مورداستفاده در این تحقیق Scikit, Pandas, Keras, Numpy و... می‌باشند (Harris et al., 2020; team, 2020). برای انجام این تحقیق ابتدا دیتابیس مربوطه که از سایت سازمان زمین‌شناسی آمریکا دانلود شده با بصری‌سازی، مورد بررسی کلی قرار گرفت. در مرحله بعد با مدل‌های متنوع یادگیری ماشین که شامل مدل خطی تک‌متغیره، خطی چندمتغیره، خطی مرتبه بالا و غیرخطی بود مورد بررسی قرارگرفت و در نهایت مدل مناسب انتخاب شد. بعد از آموزش و انتخاب مدل یادگیری ماشین مناسب، مدل شبکه عصبی آموزش داده شد که شامل یک لایه ورودی، سه‌لایه پنهانی و یک لایه خروجی است. در انتها نیز دو روش شبکه عصبی و یادگیری ماشین مورد مقایسه قرار گرفتند تا مشخص شود کدام‌یک عملکرد بهتری داشته‌اند.

یافته ها:

آماده سازی داده:

در دیتابیس موجود، بیش از ۲۰۰۰۰ سلول خالی وجود داشت که تنها حدود ۵۰۰ سلول مربوط به پارامترهای مورد استفاده در این تحقیق بودند. به دلیل تفاوت نوع آبخوان ها، زمان و مکان نمونه گیری و برخی دیگر از شرایط محیطی امکان جایگزینی داده‌ها با استفاده از مولفه ها آماری وجود نداشت و باتوجه‌به تعداد بالای ردیف داده‌های مورد استفاده، ردیف داده‌های ناقص حذف شد. در ادامه‌ی فرایند پاک‌سازی داده‌ها، تصمیم به حذف داده هایی شد که از لحاظ مفهومی تناسبی با داده‌های موردنظر نداشتند. سلول‌هایی که حاوی علائم، حروف و... بودند و ارتباطی به داده‌ها نداشتند در این گام حذف شدند. پس از این کار، داده‌های پرت با استفاده از نمودار جعبه‌ای حذف شدند. پس از پاک‌سازی و فیلتر کردن داده‌ها (بر اساس نوع آبخوان که طبق ادبیات موضوعی گفته شده) تعداد داده‌ها از ۴۸۰۰ به ۲۰۰۰ ردیف داده کاهش پیدا کرد. از ۲۰۰۰ ردیف داده موجود به شکل تصادفی ۲ دسته داده به تعداد ۱۰۰۰ ردیف انتخاب شد تا هر کدام از مدل های یادگیری ماشین و شبکه عصبی آموزش داده شوند. شکل۲ به عنوان نمونه تفاوت پراکنش و نمودار جعبه ای داده‌های مربوط به کلسیم و کل جامدات حل شده را نشان می‌دهد.

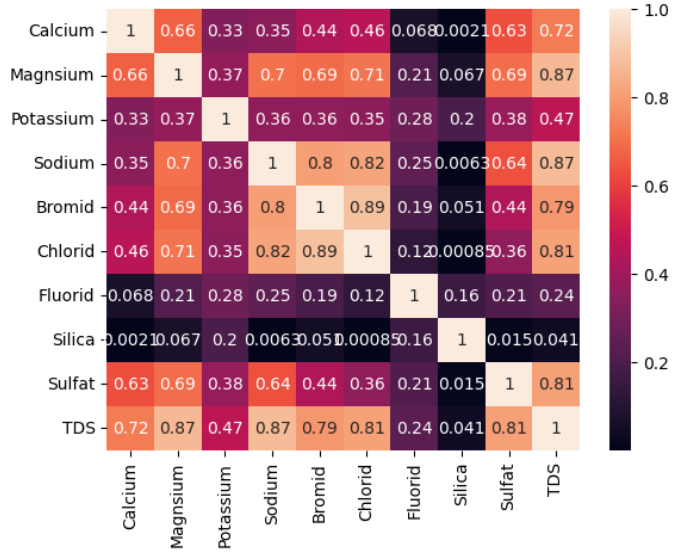




شکل ۲ نمودار پراکنش کلسیم و کل جامدات حل شده الف) قبل از آماده‌سازی داده‌ها ب)بعد از آماده‌سازی داده‌ها ج)نمودار جعبه ای کلسیم قبل از آماده‌سازی داده‌ها د) نمودار جعبه ای کلسیم بعد از آماده‌سازی داده‌ها

مدل یادگیری ماشین:

در گذشته روش های دستی و آماری متنوعی برای برازش استفاده می‌شد که در تعداد داده های زیاد بسیار سخت و وقت گیر بودند اما امروزه یکی از مزایای بسیار مهم یادگیری ماشین استفاده در مباحث مربوط به برازش است. در این مرحله از تحقیق حاضر از روش یادگیری ماشین جهت برازش یک خط در داده های موجود استفاده می‌شود. ابتدا باید پارامترهای مهم برای آموزش مدل مورد ارزیابی قرار گیرند. برای این کار از مقدار همبستگی خطی و گراف مربوط به آن استفاده می‌کنیم. نخست باتوجه‌به ادبیات موضوعی تنها عناصری که ممکن است در خروجی کل جامدات محلول مؤثر باشند را انتخاب کرده و بین این عناصر و خروجی همبستگی مورد بررسی قرار می‌گیرد. در شکل ۲-۳ مشخص است که مقادیر مربوط به کلسیم، منیزیم، سدیم، کلر و سولفات همبستگی خوبی با کل جامدات محلول دارند.



شکل ۲-۳ نمودار پراکنش کلسیم و کل جامدات حل شده الف) قبل از آماده‌سازی داده‌ها و ب)بعد از آماده‌سازی داده‌ها

در این تحقیق انواع مدل‌های خطی تک‌متغیره، خطی چندمتغیره، خطی چند درجه و غیرخطی مورد بررسی قرار گرفت و درنهایت نتایج به‌دست‌آمده طبق جدول ۱-۳ ارائه شده است. بررسی کفایت هر مدل طبق خروجی نرم‌افزار و بر اساس مقادیر ضریب تعیین می‌باشد که مدل چندمتغیره خطی با مقدار ضریب تعیین ۰.۹۳ بهترین مدل برای این داده ها می‌باشد. در این مدل کلسیم بیشترین تأثیر را داشته و کلر نیز کمترین تأثیر را داراست. جدول ۲-۳ نتایج آزمون OLS را نشان می‌دهدکه نشان‌دهنده کفایت مدل می‌باشد. در این مدل منیزیم به دلیل فراوانی در طبیعت در غلظت‌های بالا بیشترین تأثیر را روی مقدار TDS خروجی مدل داشته و کلر نیز کمترین تأثیر را دارد. برای تمام عناصر مقدار P-value از ۰.۰۵ کمتر بوده که نشان دهنده تاثیر گذاری پارامتر های انتخاب شده در مدل می‌باشد.

جدول ۱-۳ خروجی مدل یادگیری ماشین

پارامتر	مقدار
ضریب تعیین داده های آموزش	۰.۹۳
ضریب تعیین داده های آزمون	۰.۹۰
میانگین مجذورات خطا آموزش	۸۹۸
میانگین مجذورات خطا آزمون	۱۱۵۷

جدول ۲-۳ نتایج آزمون OLS

متغیر	coef	stderr	t	P> |t|	۰,۲۵}	{۰/۹۷۵
عرض از مبدأ	۳۷.۷	۲.۳	۱۶.۵	۰.۰	۳۳.۲	۴۲.۲
کلر	۰.۲	۰.۱	۲.۰	۰.۰	۰.۰	۰.۵
سدیم	۲.۴	۰.۱	۳۶.۶	۰.۰	۲.۳	۲.۶
منیزیم	۳.۲	۰.۱	۲۸.۰	۰.۰	۳.۰	۳.۴
کلسیم	۲.۰	۰.۰	۴۷.۰	۰.۰	۱.۹	۲.۱
سولفات	۰.۶	۰.۱	۱۰.۰	۰.۰	۰.۵	۰.۷

در نهایت برای جلوگیری از بیش برازش مدل سعی شد با الگوریتم Ridge تاثیرگذاری پارامتر های کم اهمیت تر را کمتر کرد. دلیل انتخاب الگوریتم Ridge عدم حذف کامل یک پارامتر توسط این الگوریتم می‌باشد زیرا در این تحقیق تمام عناصر روی مقدار خروجی مدل تاثیرگذارند. جدول ۲-۳ ضرایب هر پارامتر را پس از اصلاح نشان میدهد. پس از اصلاح مقدار ضریبی تعیین به ۰.۹۰ کاهش یافته است و میانگین مربعات خطا نیز تغییر چندانی نداشت.

جدول ۳-۳ نتایج تصحیح Ridge

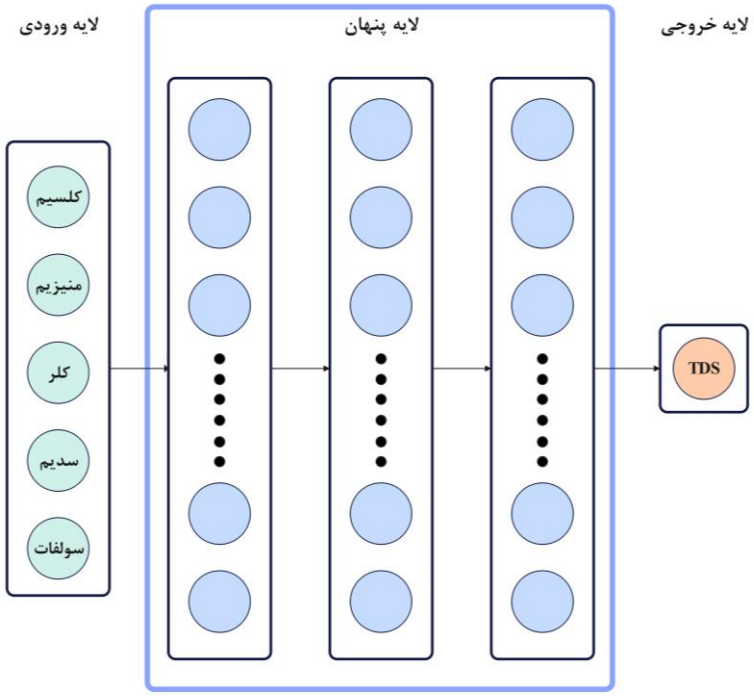
سولفات	کلسیم	منیزیم	سدیم	کلر	عرض از مبدأ

۳۹.۴	۰.۴	۲.۵	۳.۲	۲.۰	۰.۶

شبکه عصبی:

شبکه‌های عصبی مصنوعی (ANNs) ابزاری قدرتمند برای حل مسائل پیچیده در حوزه‌های مختلف از جمله علوم کامپیوتر، مهندسی، پزشکی و مالی هستند. این شبکه‌ها از ساختار مغز انسان الهام گرفته شده‌اند و از مجموعه‌ای از واحدهای پردازشی ساده به نام نورون تشکیل شده‌اند که به‌صورت متراکم به هم متصل شده‌اند. شبکه‌های عصبی پیش‌خور (FFNNs) نوعی از ANNها هستند که در آن‌ها مسیر سیگنال‌ها از ورودی به خروجی فقط در یک‌جهت است، به‌عبارت‌دیگر، هیچ حلقه بازخوردی در شبکه وجود ندارد.

در طراحی شبکه‌های عصبی به شکل کلی ۵ پارامتر تاثیرگذارند. این پارامتر ها شامل تعداد لایه، نوع لایه، اندازه لایه، توابع فعالسازی و تنظیمات نهایی میباشند. طراحی این پارامترها بسیار مهم و تأثیرگذار است و در بسیاری حالات می‌تواند پیچیده و زمانبر باشد. از این رو برای طراحی معماری شبکه عصبی از روش آزمون‌وخطا استفاده می‌شود. پس چند مرتبه آزمون‌وخطا یک معماری ۵ لایه برای شبکه عصبی انتخاب شد. در این معماری یک لایه ورودی که ۵ نورون دارد به عنوان اولین لایه در نظر گفته شده است زیرا در مدل حاضر ۵ پارامتر وجود دارد. پس وارد کرد پارامتر های ورودی سه لایه پنهانی برای پردازش اطلاعات وارده طراحی شده است. تعداد این لایه های بر اساس حجم و پیچیدگی داده انتخاب شده و با آزمون‌وخطا انجام می‌شود. پس انتخاب تعداد لایه ها، تعداد نورون موجود در هر لایه به دقت مورد ارزیابی قرار میگیرد. شکل ۳-۳ به صورت شماتیک شبکه عصبی طراحی شده را نشان می‌دهد. این مرحله از طراحی با استفاده از ابزار Keras در کتابخانه Tensorflow انجام شد. لایه های طراحی شده همه از نوع Dense بوده و تمام لایه ها به جز لایه سوم از تابع فعالسازی Relu استفاده میکنند. لایه سوم شبکه عصبی حاضر از تابع فعالسازی Softplus استفاده میکند که نسخه ای از صاف تر از تابع Relu می‌باشد. دلیل استفاده از این توابع به دلیل مثبت بودن خروجی آنها است که با توجه به فیزیک مسئله یعنی غلظت آلاینده در طبیعت منطقی به نظر میرسد.



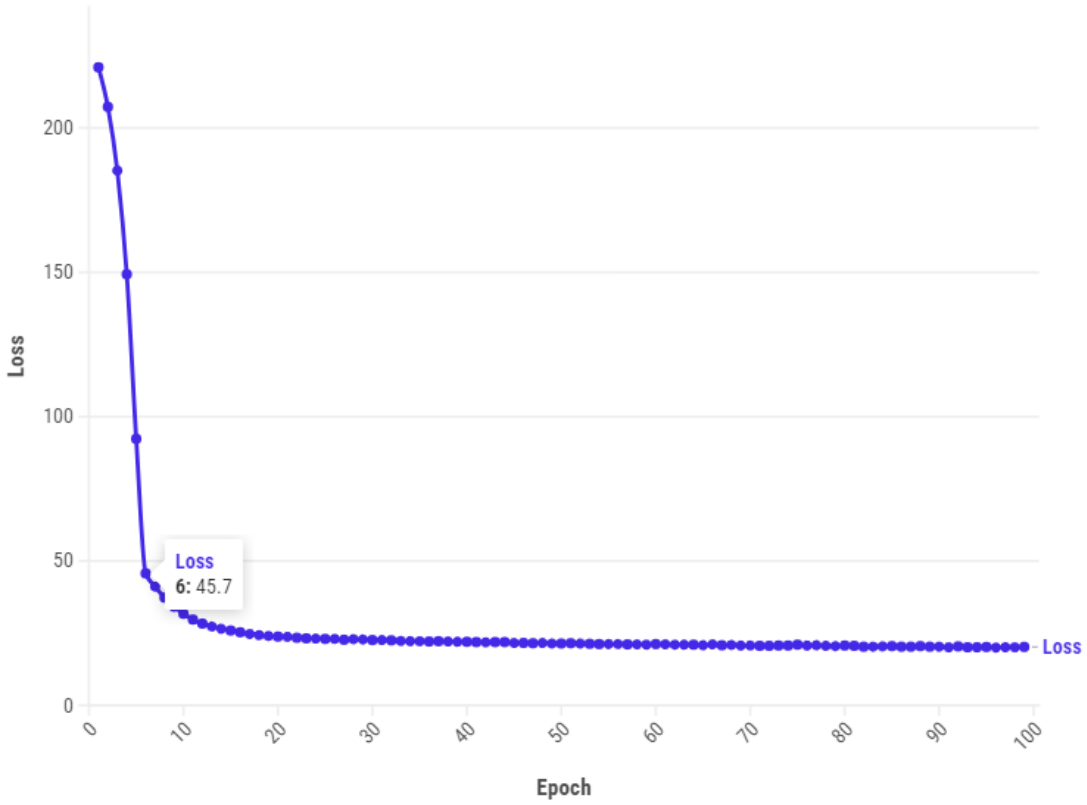
شکل ۳-۳ شماتیک شبکه عصبی مصنوعی

پس از طراحی شبکه عصبی، مدل طراحی شده روی داده‌های ورودی آموزش‌داده‌شده و سعی می‌شود بهترین تابع ممکن که بهترین عملکرد را روی داده‌های آموزش و آزمون دارد انتخاب شود. مدل خروجی در این مرحله نباید بیش برازش یا کم برازش باشد. مفاهیم بیش‌وکم برازش در این مرحله باتوجه‌به عملکرد مدل در داده‌های آموزش و آزمون تشخیص داده خواهد شد. با تغییر مقادیر مربوط به تعداد لایه‌ها و تعداد نورون‌های موجود در هر لایه مقادیر مربوط به ضریب تعیین داده‌های آزمون و آموزش تغییر می‌کند. به شکل کلی با افزایش تعداد نورون‌ها و لایه‌های موجود در مدل، بیش برازش اتفاق می‌افتد به این معنی که خط برازش داده شده به داده‌های آموزش نزدیک‌تر بوده و دقت و صحت کمتری روی داده‌های آزمون دارند. پس از درنظرگرفتن موارد فوق بهترین مدل ممکن در نظر گرفته شد و مورد ارزیابی قرار گرفت. برای برازش مدل بر داده‌های موجود، مدل در صد گام بهینه شده است.

نتایج به‌دست‌آمده از مدل مورد بررسی در جدول ۳-۳ نشان‌داده‌شده است. نتایج به‌دست‌آمده نشان می‌دهد که مدل موردنظر خوب عمل کرده و در هر دو سری داده‌های آموزش و آزمون بسیار موفق ظاهر شده است. در شکل ۴-۳ نیز سیر تغییرات خطا در هر گام بهینه‌سازی نشان‌داده‌شده است. در این شکل قابل‌مشاهده است که پس از گام ششم تغییرات بسیار کم می‌باشد؛ اما بااین‌حال روند بهینه‌سازی ادامه یافته تا بهترین مدل ممکن به دست آید.

جدول ۴-۳ خروجی مدل شبکه عصبی

پارامتر	مقدار
ضریب تعیین داده های آموزش	۰.۹۲
ضریب تعیین داده های آزمون	۰.۸۹
میانگین مجذورات خطا آموزش	۱۰۶۷
میانگین مجذورات خطا آزمون	۱۳۲۱



شکل ۴-۳ روند تغییرات خطا در هر گام بهینه‌سازی

بحث و نتیجه‌گیری:

در این تحقیق به بررسی و مقایسه روش‌های یادگیری ماشین و شبکه عصبی در پیش‌بینی مقدار کل جامدات حل شده در آب زیرزمینی پرداخته شده است. برای این کار ابتدا پاک‌سازی داده انجام شده و ردیف داده هایی که سلول حالی داشتند حذف شدند. سپس داده‌های موجود به شکل تصادفی به دو دسته مجزا برای استفاده در مدل یادگیری ماشین و شبکه عصبی تقسیم شد. هرکدام از این دسته‌ها نیز با نسبت ۸۰ به ۲۰ به دو دسته آموزش و آزمون تقسیم شدند. در مرحله بعد یک مدل یادگیری ماشین رگرسیون ساخته شده و روی داده‌های آموزش یادگیری ماشین آموزش داده شد. سپس مدل به‌دست‌آمده از لحاظ کفایت و بیش یا کم برازش بودن مورد بررسی قرار گرفت. در این مرحله مدل خطی چندگانه بهترین خروجی را ارائه داد. پس از آزمودن مدل یادگیری ماشین، مراحل ساخت و آموزش مدل شبکه عصبی مصنوعی انجام شد. در این مرحله با آزمون‌وخطا یک مدل ۵ لایه طراحی شد و روی داده‌های آموزش شبکه عصبی آموزش داده شد. در انتها نیز بیش یا کم برازش بودن مدل شبکه عصبی مورد بررسی قرار گرفت. در نتایج این تحقیق دیده شد که هر دو مدل یادگیری ماشین و شبکه عصبی مصنوعی عملکرد تقریبا مشابهی داشتند. روش یادگیری ماشین با ضریب تعیین ۰.۹۰ در داده های آزمون عملکرد بهتری نسبت به روش شبکه عصبی داشت که ضریب تعیین ۰.۸۸ را در خروجی مدل ارائه کرد. مقدار میانگین مجذورات خطا در مدل یادگیری ماشین به ترتیب ۱۵ و ۱۲ درصد در داده های آموزش و آزمون کمتر بود. در انتها مدل یادگیری ماشین از لحاظ مصرف منابع پردازشی بهینه‌تر عمل کرد و جواب نهایی تقریباً ۸ برابر سریع‌تر به دست آمد. به شکل کلی در مواردی که مدل پیچیدگی زیادی دارد و اندرکنش چند متغیر در نتیجه نهایی مؤثر است، استفاده از مدل‌های شبکه عصبی می‌تواند راه‌حل بهتری باشد؛ اما در شرایطی که رابطهٔ بین متغیرهای مستقل و وابسته دارای پیچیدگی زیادی نیست، استفاده از مدل یادگیری ماشین می‌تواند عملکرد بهتری داشته و باعث صرفه‌جویی در مصرف منابع پردازشی شود.

منابع:

عزیزپور، علی، ایزدبخش، محمدعلی، شعبانلو، سعید، یوسفوند، فریبرز، و رجبی، احمد. (۱۴۰۰). شبیه سازی تراز، کلر و بی کربنات آب زیرزمینی توسط ماشین آموزش ترکیبی. هیدروژئولوژی، ۶(۱)، ۹۹-۱۱۳ https://sid.ir/paper/1036292/fa . SID.

ترابی پوده، حسن، نصرالهی، علی حیدر، و دهقانی، رضا. (۱۴۰۰). ارزیابی مدل شبکه عصبی موجک در پیش بینی منابع آب زیرزمینی (مطالعه موردی: استان لرستان, ایران). هیدروژئولوژی، ۶(۱)، ۱-۱۲ https://sid.ir/paper/1036278/fa . SID.

دانشوروثوقی، فرناز، و منافیان آذر، وحید. (۱۳۹۷). استفاده از مدل های ترکیبی ماشین بردار پشتیبان-موجکی و شبکه عصبی-موجکی در پیش بینی تراز آب زیرزمینی دشت اردبیل. هیدروژئومورفولوژی، ۵(۱۷)، ۴۵-۶۴ https://sid.ir/paper/388620/fa . SID.

دانشوروثوقی، فرناز، و کریمی، علی. (۱۳۹۷). استفاده از روش های پیش پردازش SOM و تبدیل موجک در پیش بینی تراز آب زیرزمینی (مطالعه موردی: دشت آذرشهر). هیدروژئولوژی، ۳(۱)، ۱۵-۳۱. https://sid.ir/paper/268154/fa SID.

Amjad Aliewi, J. A.-K., Asim Al-Khalid, Harish Bhandary & Habib Al-Qallaf (2020). Modelling the effect of high level of total dissolved solids (TDS) for the sustainable utilization of brackish groundwater from saline aquifers in Kuwait. *Environment, Development and Sustainability*. https://doi.org/https://doi.org/10.1007/s10668-020-00670-9

Freeze R.A. , C. J. A. (1979). *Groundwater*. Prentice-Hall Inc .

Ghaeminia, M., & Mokhtarani, N. (2018). Remediation of nitrate-contaminated groundwater by PRB-Electrokinetic integrated process. *J Environ Manage*, *222*, 234-241. https://doi.org/10.1016/j.jenvman.2018.05.078

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Rio, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Musgrove ,M. (2020). *Data for the occurrence and distribution of strontium in U.S. groundwater* U.S. Geological Survey - ScienceBase. https://doi.org/https://doi.org/10.5066/P9XV6DZO

Nejatijahromi, Z., Nassery, H. R., Hosono, T., Nakhaei, M., Alijani, F., & Okumura, A. (2019). Groundwater nitrate contamination in an area using urban wastewaters for agricultural irrigation under arid climate condition, southeast of Tehran, Iran. *Agricultural Water Management*, *221*, 397-414. https://doi.org/10.1016/j.agwat.2019.04 ۰۱۵.

Priyanka, P., Krishan, G., Sharma, L. M., Yadav, B., & Ghosh, N. C. (2016). Analysis of Water Level Fluctuations and TDS Variations in the Groundwater at Mewat (Nuh) District, Haryana (India). *Current World Environment*, *11*(2), 388-398. https://doi.org/10.12944/cwe.11.2.06

Programme, U. W. W. A. (2022). *GROUNDWATER: Making the invisible visible* (S. a. C. O. the United Nations Educational, Ed.) .

team, T. p. d. (2020). *pandas-dev/pandas: Pandas*. In Zenodo .

Yujie Ji, J. W., Yuanhang Wang, Vetrimurugan Elumalai & Thirumalaisamy Subramani (2020). Seasonal Variation of Drinking Water Quality and Human Health Risk Assessment in Hancheng City of Guanzhong Plain, China. *Exposure and Health*. https://doi.org/10.1007/s12403-020-00357-6

Zhang, S., Mao, G ,.Crittenden, J., Liu, X., & Du, H. (2017). Groundwater remediation from the past to the future: A bibliometric analysis. *Water Res*, *119*, 114-125. https://doi.org/10.1016/j.watres.2017.01.029

Zhou Xun, Z. H., Zhao Liang, Shen Ye, Yan Xia, Li Rui & Zhang Li (2007). Some factors affecting TDS and pH values in groundwater of the Beihai coastal area in southern Guangxi, China. *Environmental Geology* .

Prediction of Total Dissolved Solids in Groundwater Using Machine Learning and Artificial Neural Networks

Mobin Vakili
Amirkabir University of Technology

Taghi Ebadi
Amirkabir University of Technology

Amir Golroo
Amirkabir University of Technology

Abstract

Recent years have witnessed a concerning decline in the quality and quantity of groundwater, the largest readily available source of fresh water for humans. This problem has led to the exploration of various remediation techniques, such as electrokinetics, for water restoration. Designing such methods necessitates specific aquifer characteristics, including electrical conductivity (EC). Since EC determination requires total dissolved solids (TDS) concentration, this study investigated the application of machine learning and artificial neural networks (ANNs) for predicting TDS in groundwater. A comprehensive database of US groundwater characteristics spanning 1991-2018 was employed for model development. The machine learning model employed multiple linear regression, while the artificial neural network architecture consisted of five hidden layers of forward-feeding neural networks. The analysis revealed comparable performance between both methods, with a coefficient of determination (R^2) of 0.93 for test data. The investigation also encompassed the phenomena of overfitting and underfitting. However, performance reports indicated that the machine learning approach exhibited superior computational resource management, requiring approximately eight times fewer resources compared to the ANN model.

Keywords: Groundwater – Machine Learning – Data Science – Artificial Neural Networks- Total Disolved Solids- Water Quality