



Available online: <http://ijhe.tums.ac.ir>

مقاله پژوهشی



کاربرد تکنیک داده‌کاوی درخت تصمیم CART در تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی (مطالعه موردی: دشت کازرون استان فارس)

سید مسعود سلیمان‌پور^{۱*}، سید حمید مصباح^۱، بهرام هدایتی^۲

۱- بخش تحقیقات حفاظت خاک و آبخیزداری، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان فارس، سازمان تحقیقات، آموزش و ترویج کشاورزی، شیراز، ایران

۲- گروه مهندسی کامپیوتر، دانشگاه پیام نور، تهران، ایران

اطلاعات مقاله: چکیده

زمینه و هدف: بررسی پارامترهای کیفیت آب آشامیدنی به منظور افزایش بهره‌وری و مدیریت و برنامه‌ریزی بهتر منابع آب به‌خصوص در کشورهای در حال توسعه از اهمیت شایانی برخوردار است. هدف از انجام پژوهش حاضر، کاربرد تکنیک داده‌کاوی درخت تصمیم CART در تعیین فاکتورهای مؤثر بر کیفیت آب آشامیدنی در دشت کازرون واقع در غرب استان فارس است.

روش بررسی: به این منظور، اطلاعات مربوط به پارامترهای کیفی ۶۰ حلقه چاه شرب، شامل SAR، EC، Na، Cl، SO₄، TH، TDS، pH، NO₃، CaCO₃، Ca، Mg، K و EC، مستقر در منطقه مورد مطالعه فراهم گردید. سپس، با استفاده از تکنیک داده‌کاوی درخت تصمیم CART و مدل‌سازی در نرم افزار Clementine 12.0، فاکتورهای مؤثر بر کیفیت آب آشامیدنی با دقت ۹۰ درصد تعیین گردیدند.

یافته‌ها: نتایج نشان داد که دو عامل کل جامدات محلول (TDS) و مقدار کلسیم (Ca)، تاثیر بیشتری بر کیفیت آب آشامیدنی در این دشت داشته است. به این ترتیب، در صورتی که مجموع املاح محلول در آب در این دشت کمتر یا مساوی ۴۹۵ mg/L و مقدار کلسیم آن، کمتر یا مساوی ۶/۱۵۰ meq/L باشد، این آب برای آشامیدن مناسب است.

نتیجه‌گیری: با توجه به ساختار سازندهای زمین‌شناسی منطقه و وجود کربنات کلسیم در ترکیب آنها، کل جامدات محلول و کلسیم تاثیر بیشتری بر کیفیت آب آشامیدنی در این دشت دارند. پیشنهاد می‌گردد نسبت به تصفیه آب، کاهش املاح محلول و انجام پایش‌های مستمر از چاه‌های این دشت اقدام صورت گیرد.

تاریخ دریافت: ۹۶/۰۹/۰۷

تاریخ ویرایش: ۹۶/۱۱/۲۸

تاریخ پذیرش: ۹۶/۱۲/۰۲

تاریخ انتشار: ۹۷/۰۳/۳۰

واژگان کلیدی: آب آشامیدنی، داده‌کاوی، درخت تصمیم، دشت کازرون، کیفیت آب

پست الکترونیکی نویسنده مسئول:

m.soleimanpour@yahoo.com

مقدمه

در کشورهای غربی و صنعتی سالانه هر فرد حداقل میزان به 2000 m^3 آب برای برخورداری از یک استاندارد مطلوب نیاز دارد. اگر سرانه آب هر فرد بین 1000 تا 2000 m^3 باشد آن کشور تحت تنش آب است. اما در صورتی که سرانه از 500 m^3 در سال کمتر باشد، کشور مذکور با کم‌آبی مواجه است. در حال حاضر منابع آب موجود می‌تواند سالانه 7000 m^3 برای هر فرد آب فراهم نماید. اگرچه آب کافی حداقل برای سه برابر جمعیت کره زمین موجود است، ولی عدم تعادل بین توزیع جمعیت و بارندگی، موجب کمبود آب و استفاده از آب‌های آلوده در بعضی از مناطق شده است (۱)؛ از این‌رو، وجود آب آشامیدنی سالم، ضامن سلامتی جامعه است و اولین قدم در دستیابی به این مهم، بررسی پارامترهای آن است (۲). علاوه بر این، مهمترین مسائلی که در اثر مصرف آب‌های نامناسب در امور کشاورزی ایجاد می‌شود شامل شور شدن ثانویه خاک‌ها، کاهش نفوذپذیری و سمیت املاح است؛ که هر کدام به نوعی بر سلامتی و یا رشد و تولید محصولات (کشاورزی، دامداری، صنعتی و ...) و در نهایت آسیب به سیستم ایمنی و سلامت انسان خدشه وارد می‌نماید (۳).

وضعیت منابع آب شرب (به‌ویژه منابع شرب سالم)، در جهان روز به روز در حال کاهش است. سازمان خواربار و کشاورزی ملل متحد، علاوه بر اعلام نگرانی در این خصوص، پیش‌بینی کرده دو سوم جمعیت جهان در سال 2025 میلادی با کمبود آب روبرو خواهند بود (۴).

ایران، کشوری پهناور است که از نظر منابع آبی محدود است. به همین جهت از زمان‌های دور، آب در کشور ما از ارزش و اهمیت والایی برخوردار بوده است. شرایط اقلیمی، زمین‌شناسی و آب‌شناسی کشور ما نیز طوری است که بهره‌برداری از آب با کیفیت را با مسائل و مشکلات خاص خود همراه می‌کند. انسان برای آشامیدن، بهداشت فردی و اجتماعی، پخت و پز، شست و شو و آبیاری به آب مناسب و با کیفیت احتیاج دارد و هر انسان در روز حداقل به $2/5 \text{ L}$ آب شیرین سالم و با کیفیت استاندارد برای آشامیدن نیاز دارد تا بتواند احتیاجات خود را

برطرف نماید. علاوه بر آن، برای هر نفر در کشور ایران به‌طور متوسط روزانه به 110 L آب برای مصارف شخصی و خانگی نیاز است که در واقع همان سرانه مصرف آب در شهرهای کشور است (۵). به این منظور و در پی کمبود جهانی آب، پیش‌بینی کمیت و کیفیت آب، توجه جدی پژوهشگران هیدرولوژی را به خود جلب کرده است (۶). این امر به ویژه در شرایط خشک و نیمه‌خشک کشور و کمبود منابع آب شیرین، حساسیت نسبت به کیفیت آب و عوامل مؤثر بر آنها را ضروری‌تر نموده است (۷)؛ بنابراین بررسی وضعیت پارامترهای شیمیایی در آب نمی‌تواند به تنهایی گویای وضعیت فعلی آب باشد و برای تشخیص کیفیت آب بایستی توجه بیشتری به شاخص‌های کیفی آب معطوف نمود و کاربرد این شاخص‌ها را به صورت یک معیار ضروری جهت ارزیابی کیفیت آب شرب در شبکه‌های توزیع و منابع تامین آب مورد توجه قرار داد (۸). در نتیجه بررسی پارامترهای کیفیت آب به منظور افزایش بهره‌وری و توسعه مدیریت و برنامه‌ریزی بهتر منابع، امری ضروری است (۹)؛ به این منظور یکی از راه‌های اثربخش و تاثیرگذار برای مبارزه با این معضل ملی، استفاده بهینه از منابع آبی کشور با تاکید بر اصول کیفیت آب است، به‌طوری که امروزه بررسی‌های کیفی آب، دامنه گسترده‌ای پیدا کرده و مسائل مربوط به آلودگی آب‌های سطحی و زیرزمینی را نیز شامل می‌شود. این مبحث نه تنها در کشورهای صنعتی، بلکه در کشورهای در حال توسعه نیز مطرح است (۱۰).

داده‌کاوی یک هنر و علم استخراج اطلاعات پنهان، از مجموعه داده‌های فراوان است (۱۱). این علم، به‌عنوان مجموعه‌ای از روش‌ها در فرایند کشف دانش است که برای تشخیص الگوها و رابطه‌های نامعلوم در داده‌ها مورد استفاده قرار می‌گیرد (۱۲). به‌عبارت دیگر، داده‌کاوی مجموعه‌ای از فعالیت‌هایی است که برای یافتن الگوهای جدید، پنهان، و غیر منتظره در داده‌ها استفاده می‌شود (۱۳).

از دلایل رشد و توسعه دانش داده‌کاوی در علوم مختلف، می‌توان به افزایش حجم پایگاه داده‌ها و عدم توانایی انسان برای درک و استخراج از آنها اشاره کرد. بنابراین درک این

کمتر یا مساوی $731/65 \mu\text{mhos/cm}$ باشد، این آب مناسب آشامیدن است. در پژوهش دیگری Arora و همکاران (۱۶) با استفاده از تکنیک‌های داده‌کاوی و خوشه‌بندی، اقدام به ارزیابی کیفیت آب رودخانه ساتلوج واقع در کشور هند کردند. ایشان روش‌های مذکور را مناسب و قابل اطمینان معرفی، و استفاده از آنان را به مدیران منابع آب توصیه نمودند. به‌علاوه Cho (۱۷) در پژوهشی به ارزیابی مدل کیفیت آب در حوزه‌های آبخیز با استفاده از داده‌کاوی پرداخت. ایشان در این تحقیق از تکنیک‌های درخت تصمیم، شبکه عصبی مصنوعی، و توابع شعاعی استفاده نمود و به این نتیجه رسید که مدل ارزیابی کیفیت آب براساس داده‌کاوی، جایگزین مناسبی برای مدل‌های فیزیکی در حوزه‌های آبخیز است. همچنین Soleimanpour و همکاران (۵) در پژوهشی اقدام به تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی با استفاده از تکنیک‌های داده‌کاوی QUEST (Quick Unbiased and Efficient Statistical Tree) در شهرستان سعادت‌شهر استان فارس نمودند. نتایج نشان داد مؤثرترین فاکتورهای کیفیت آب آشامیدنی در این منطقه تابع سختی کل و هدایت الکتریکی است. به‌علاوه Rajakumari و همکار (۱۱) اقدام به مقایسه روش‌های مختلف خوشه‌بندی به‌منظور بررسی آلاینده سرب در آب رودخانه‌ها کردند. در این پژوهش از روش‌های خوشه‌بندی K-Means، EM (Expectation Maximization)، سلسله مراتبی و تار عنکبوت استفاده نمودند و به این نتیجه رسیدند که از بین روش‌های مختلف خوشه‌بندی برای این منظور، روش K-Means دارای مقبولیت بیشتری است. همچنین Ji و همکاران (۱۸) در پژوهشی با استفاده از تکنیک‌های داده‌کاوی اقدام به طبقه‌بندی و برنامه‌ریزی مخازن تامین آب نمودند. ایشان از شبکه عصبی (Multi Layer) ML و درخت تصمیم CART استفاده و با استفاده از برنامه‌نویسی پویای DP (Dynamic Programming)، به این نتیجه رسیدند که استفاده از روش درخت تصمیم CART، نتایج بهتری در این خصوص ارائه می‌دهد. Fu-Cheng و همکار (۱۹) از روش خوشه‌بندی فازی

داده‌ها بدون ابزارهای قدرتمند میسر نیست؛ زیرا تصمیم‌گیران ابزار قوی برای استخراج اطلاعات با ارزش را در دست ندارند. در واقع شرایط فعلی توصیف‌کننده حالتی است که ما از لحاظ داده غنی، اما از لحاظ اطلاعات ضعیف هستیم؛ حال با توجه به شدت رقابت‌ها در عرصه‌های مختلف، استفاده مؤثر از داده‌ها توسط مدیران، یک هدف عمده برای بهبود وضعیت موجود محسوب می‌شود (۱۴). به این منظور امروزه بسیاری از سازمان‌ها از تکنیک‌های داده‌کاوی استفاده می‌کنند و تحقیقات در زمینه داده‌کاوی در حال رشد در تمام زمینه‌ها در طول دهه آینده است (۱۳).

درخت تصمیم، یکی از مشهورترین تکنیک‌های دسته‌بندی است که در فرایند داده‌کاوی کاربرد دارد. شیوه‌نمایش درخت تصمیم به این صورت است که با ارائه نمودن یک درخت، روال دسته‌بندی را خلاصه‌سازی می‌کند. از درخت‌های تصمیم‌گیری به‌منظور پیش‌بینی کردن عضویت اشیاء به دسته‌های مختلف استفاده می‌شود. انعطاف‌پذیری این تکنیک باعث شده تا در میان روش‌های جذاب داده‌کاوی، بیشتر مورد استفاده قرار گیرد. متدولوژی مربوط به درخت تصمیم شامل دو فاز اصلی است:

الف- ساخت درخت اولیه: با استفاده از مجموعه داده‌های آموزشی، ساخت درخت تصمیم تا زمانی که هر برگ، خالص (همگن) شود ادامه می‌یابد.

ب- هرس کردن: در این فاز، به‌منظور افزایش دقت مدل، درخت رشد یافته با توجه به مجموعه داده‌های آزمایشی هرس می‌شود (۱۵).

Soleimanpour و همکاران (۱۰) نسبت به کاربرد الگوریتم‌های داده‌کاوی K-Means و CART (Classification And Regression Trees) در تعیین مؤثرترین عوامل کیفیت آب آشامیدنی در دشت نورآباد استان فارس اقدام کردند. نتایج این تحقیق نشان داد مؤثرترین فاکتورهای مطلوب آب آشامیدنی در این منطقه، تابع سختی کل و هدایت الکتریکی عصاره اشباع است. بدین ترتیب، در صورتی که سختی کل در این دشت بین ۱۷۵ و ۴۴۰ ppm، و هدایت الکتریکی عصاره اشباع آن،

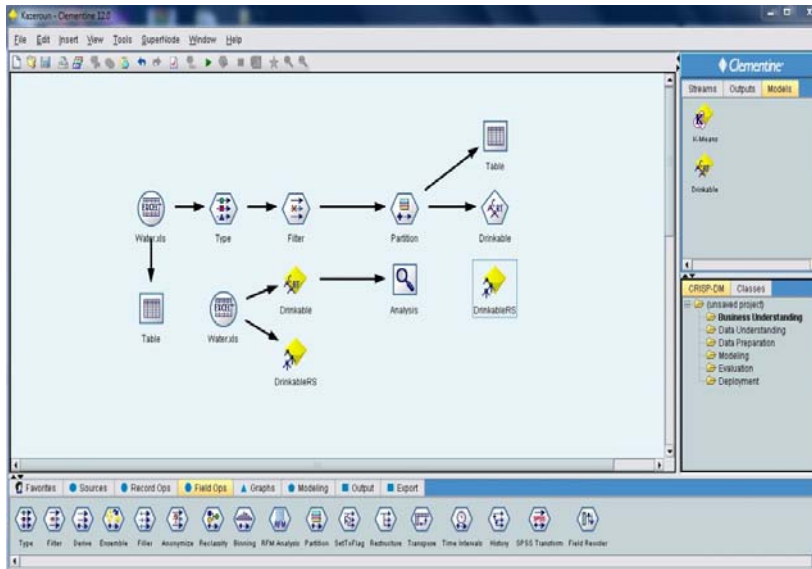
بهبتر مدیریت کردن آنها اقدامات مؤثری را به عمل آورد. بدین منظور پژوهش حاضر نسبت به کاربرد تکنیک داده‌کاوی درخت تصمیم CART در تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی در دشت کازرون واقع در غرب استان فارس اقدام نموده است.

مواد و روش‌ها

به‌منظور انجام این پژوهش، با مراجعه به سازمان آب منطقه‌ای استان فارس، نسبت به تهیه پارامترهای کیفی چاه‌های شرب مستقر در دشت کازرون واقع در غرب استان فارس اقدام شد و مشخص گردید که در این دشت تعداد ۷۰ حلقه چاه دارای آمار کامل پارامترهای کیفی آب بود. به این منظور جهت تعیین تعداد اصولی نمونه به کمک فرمول کوکران (Cochran) و با خطای ۵ درصد، تعداد ۶۰ حلقه چاه انتخاب شد. محدوده مطالعاتی در موقعیت جغرافیایی $51^{\circ}39'15''$ طول شرقی، و $29^{\circ}27'$ عرض شمالی واقع شده است. آمارهای اخذ شده برای ۶۰ حلقه چاه، مشتمل بر فاکتورهای کیفی SAR, Na, Cl, SO₄, TH, EC, K, Mg, Ca, HCO₃, CaCO₃, NO₃, pH, TDS بود. به‌منظور نرمال بودن داده‌ها، از نرم افزار SPSS 16، و آزمون کولموگوروف-اسمیرنوف (Kolmogorov-Smirnov) استفاده شد و نرمال بودن داده‌ها تایید شد. سپس به‌منظور تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی از تکنیک داده‌کاوی درخت تصمیم CART استفاده شد. مدل‌سازی در نرم افزار Clementine 12.0 انجام گرفت. نرم افزار یاد شده ساخت شرکت SPSS است و امکان ایجاد مدل‌های متعددی را براساس تئوری‌های آماری، هوش مصنوعی و یادگیری ماشین ارائه می‌دهد. شکل ۱ نمایی از مدل‌سازی انجام شده در پژوهش حاضر را با استفاده از این نرم افزار نشان می‌دهد. در این نرم افزار، کلیه فرایندها به صورت جریان (Stream) طراحی می‌شوند. یک جریان شامل مجموعه‌ای از گره‌ها (Node) است که به ترتیب به یکدیگر متصل می‌شوند؛ به گونه‌ای که خروجی هر گره، ورودی گره بعدی باشد تا در نهایت هدفی را برآورده سازند. هر گره در جریان، وظیفه‌ای بر عهده دارد و مانند یک

c-means به‌منظور طبقه‌بندی و ارزیابی کیفیت آب‌های سطحی روستایی ۳۳ ایستگاه در شهر لیان یانگانگ استفاده و مناطق آلوده به آمونیاک و نیتروژن را تعیین کردند. ایشان کاربرد این روش را در شناسایی مناطق آلوده و ارزیابی کیفی منابع آب، مطلوب ارزیابی نمودند. Azhar و همکاران (۲۰) هم از منطق فازی و درخت تصمیم به صورت توأم در بررسی استانداردهای کیفیت آب استفاده کردند و اعلام داشتند به‌منظور نظارت بر کیفیت آب، بهره‌گیری از این روش‌ها مناسب است. به‌علاوه Gu و همکاران (۲۱) در پژوهشی به شناسایی و ارزیابی عوامل تاثیرگذار بر کیفیت آب در ۷۳ مخزن آب آشامیدنی در استان ژجیانگ چین اقدام کردند. نتایج این تحقیق که با بهره‌گیری از درخت تصمیم‌گیری CART، به‌دست آمد بیانگر ۸۱ درصد دقت در تخمین کیفیت آب بود. با توجه به نتایج مطلوب، ایشان استفاده از این روش را در مطالعه کیفیت آب آشامیدنی سایر مناطق پیشنهاد و بیان کردند که این روش می‌تواند به‌طور بالقوه به‌عنوان ابزار عملیاتی برای برنامه‌ریزان و مدیران عمل نماید. همچنین Lee و همکار (۲۲) در تحقیقی پتانسیل بهره‌وری آب زیرزمینی را با استفاده از تکنیک درخت تصمیم و سیستم اطلاعات جغرافیایی در شهرهای بورایونگ و پوهانگ کره، مورد بررسی قرار دادند. نتایج نشان داد که مدل‌های درخت تصمیم‌گیری می‌توانند برای مطالعه و توسعه منابع آب زیرزمینی مفید باشند. Thompson (۲۳) نیز در پژوهشی با استفاده از تکنیک‌های درخت تصمیم به بررسی کیفیت آب آشامیدنی در کانادا پرداخت و اعلام نمود نتایج حاصل از روش‌های داده‌کاوی و درخت تصمیم می‌تواند در اطلاع‌رسانی، و تصمیم‌گیری‌های آینده مفید واقع گردد.

با عنایت به مباحث فوق، باید اذعان نمود که امروزه بحث کیفیت آب در بسیاری از مناطق جهان به‌عنوان یکی از مباحث کلیدی مطرح است. زیرا این امر ارتباط بسیاری با سلامتی انسان و جامعه بشری دارد. به طوری که بررسی کیفیت آب (به ویژه آب آشامیدنی)، و چگونگی تغییر آنها نقش بسیار مهمی در مدیریت و بهره‌برداری از منابع دارد. به‌علاوه با آگاهی از فاکتورهای مؤثر در تغییر کیفیت آب آشامیدنی می‌توان در جهت هر چه



شکل ۱- مدلسازی انجام شده با استفاده از نرم افزار SPSS Clementine 12.0

شامل اطلاعات زیادی باشد و اصولاً تعدادی از فیلدهای آن در موضوع مورد تحقیق قابل استفاده نباشند، می‌توان با استفاده از گره **Filter** از ورود ستون‌های غیر ضروری به بخش‌های بعدی جریان جلوگیری نمود.

گره **Partition** که تنظیمات آن در شکل ۳ نشان داده شده است به منظور مشخص کردن درصد داده‌های آموزشی و آزمایشی جهت تقسیم کردن است که توضیحات کامل آن در ادامه و در بخش دسته‌بندی ارائه شده است.

گره **CART** که در شکل ۱، پس از گره **Partition** قرار گرفته است، الگوریتم درخت تصمیم **CART** را بر روی داده‌های ورودی به خود اعمال و نتایج را در قالب یک مدل تولید می‌نماید (توضیحات این الگوریتم در بخش درخت تصمیم **CART** ارائه شده است).

مدل تولید شده به رنگ زرد است و پس از تولید شدن در بخش **Models** قابل مشاهده است (مدل تولید شده و الگوهای استخراج شده توسط آن در شکل ۴ نشان داده شده‌اند). به‌منظور اعتبارسنجی مدل تولید شده و به‌دست آوردن میزان اطمینان آن، بایستی مدل تولید شده از پنل مذکور به صفحه اصلی برنامه منتقل شود و خروجی آن وارد گره **Analysis**

دستگاه عمل می‌کند؛ یعنی مجموعه‌ای از ورودی‌ها را می‌پذیرد و یک یا چند خروجی تولید می‌نماید.

انواع گره‌هایی که در شکل فوق استفاده شده‌اند به همراه روند طی شدن جریان، به شرح زیر است:

گره **Excel** (نخستین گرهی که جریان از آنجا آغاز شده است) یک گره منبع محسوب می‌شود و با دریافت آدرس یک فایل اکسل، اطلاعات ثبت شده در آن فایل را برای استفاده‌های بعدی وارد نرم افزار می‌کند.

از گره **Table** جهت نمایش اطلاعات ورودی در قالب یک جدول استفاده می‌شود.

با استفاده از گره **Type** می‌توان نوع داده مورد استفاده برای هر ستون از اطلاعات ورودی را تعیین کرد. به‌علاوه، جهت تعیین متغیر هدف از این گره استفاده می‌شود. برای مثال مقادیر موجود در ستون **Ca** از نوع عددی و مقادیر ثبت شده در ستون **Drinkable** از نوع متنی هستند. همچنین، فیلد **Drinkable** به دلیل مشخص کردن قابل آشامیدن یا غیر قابل آشامیدن بودن آب هر رکورد اطلاعاتی، به‌عنوان متغیر هدف نیز تعیین شده است.

با توجه به این که یک مجموعه داده (**Dataset**)، ممکن است

تقسیم می‌شوند (۲۴). همان‌طور که در شکل ۲ نشان داده شده است، الگوریتم‌های دسته‌بندی شامل دو مرحله‌ی آموزش و آزمایش هستند. در مرحله آموزش، الگوریتم یادگیرنده براساس مجموعه داده‌های آموزشی، یک مدل را تولید می‌کند (۲۵). بنابراین دسته‌بندی از جمله روش‌هایی در داده‌کاوی هست که در آن برای هر کدام از رکوردهای مجموعه داده‌های مورد کاوش، یک برچسب که بیانگر حقیقتی در مساله است، وجود دارد. این برچسب سبب می‌شود که هر الگوریتم دسته‌بندی، یک الگوریتم با نظارت محسوب گردد. در روش‌های با نظارت، الگوریتم ابتدا در مرحله آموزش مدلی را فرا می‌گیرد و سپس در مرحله ارزیابی، کارایی مدلی یاد گرفته بررسی می‌شود. این الگوریتم‌ها جزء الگوریتم‌های با نظارت هستند زیرا هر رکورد در مجموعه داده‌های آموزشی و آزمایشی، دارای برچسبی مشخص است و هدف الگوریتم، یادگیری نظم حاکم بر انواع برچسب‌ها (که در این تحقیق دو نوع برچسب وجود دارند: آشامیدنی، غیر قابل آشامیدن) براساس سایر ویژگی‌های رکوردها است. در

گردد. این گره براساس داده‌های آزمایشی مدل تولید شده (الگوهای کشف شده) را مورد ارزیابی قرار می‌دهد و درصد رکوردهایی که به‌طور صحیح پیش‌بینی شده‌اند را نمایش می‌دهد. پیش‌بینی، بر مبنای متغیر هدف مشخص شده (Drinkable) انجام می‌پذیرد.

با عنایت به تشریح مراحل مدلسازی، و با توجه به این که در مبحث داده‌کاوی، مهمترین موضوع دستیابی به داده‌هایی است که بتوان براساس آنها به نتایج مفیدی دست یافت. در این پژوهش، ۱۴ فاکتور (ویژگی) کیفی ۶۰ حلقه چاه آب موجود در منطقه به‌عنوان ورودی الگوریتم CART تعیین گردیدند. جدول ۱، فاکتورهای کیفی گروه‌بندی شده مناسب جهت شرب انسان بر مبنای طبقه‌بندی‌های ویلکوکس، شولر، و اوکین و استاندارد کیفیت آب آشامیدنی ایران را نشان می‌دهد.

مراحل فرایند دسته‌بندی

در الگوریتم‌های دسته‌بندی، کل مجموعه داده‌ها به دو قسمت مجموعه داده‌های آموزشی و مجموعه داده‌های آزمایشی

جدول ۱- فاکتورهای کیفی آب آشامیدنی

فاکتورهای کیفی و محدوده‌های مجاز	واحد
$\text{CaCO}_3 < 500$	meq/L
$\text{HCO}_3 < 500$	meq/L
$\text{NO}_3 < 50$	meq/L
$\text{SO}_4 < 400$	meq/L
$\text{Ca} < 300$	meq/L
$\text{Cl} < 400$	meq/L
$\text{K} < 12$	meq/L
$\text{Mg} < 150$	meq/L
$\text{Na} < 200$	meq/L
$\text{EC} < 750$	$\mu\text{mhos/cm}$
$\text{TDS} < 1500$	mg/L
$\text{TH} < 500$	ppm
$6/5 < \text{pH} < 9$	-
$\text{SAR} < 18$	-

روش‌های درخت تصمیم به‌ویژه در آشکار کردن روابط پیچیده بین متغیرها، بسیار توانمند هستند. هر شاخه‌ای از درخت می‌تواند شامل ترکیبات مختلفی از متغیرها باشد و متغیرهای یکسان می‌توانند بیش از یک بار در قسمت‌های مختلف درخت ظاهر شوند. این امر می‌تواند مشخص کند که چگونه یک متغیر می‌تواند به متغیر دیگری وابسته باشد (۲۶).

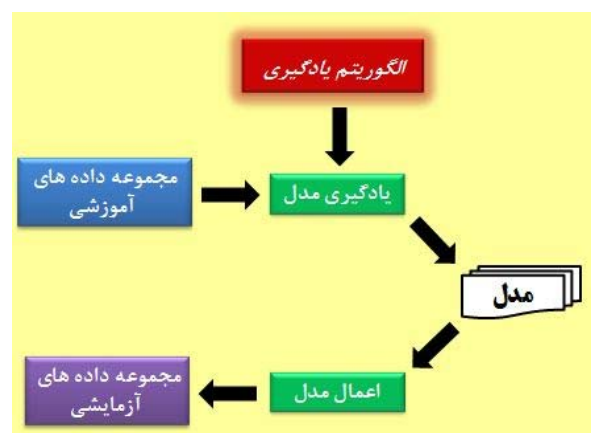
درخت تصمیم CART

درخت‌های دسته‌بندی و رگرسیون در سال ۱۹۸۴ توسط Braiman و همکاران معرفی شدند. ایده اصلی این روش، تقسیم داده‌ها به بخش‌های کوچکتر است به طوری که این بخش‌ها حاوی اطلاعات تا حد امکان تفکیک شده باشند (۲۷). الگوریتم درخت تصمیم CART روشی برای ساخت مدل‌های پیش‌بینی از داده‌ها است. این الگوریتم داده‌های ورودی به خود را به‌صورت بازگشتی تقسیم می‌نماید و قادر به پردازش متغیرهای پیشگو و هدف عددی و دسته‌ای است (۲۸). به‌منظور استفاده از الگوریتم CART، بایستی ابتدا یک ویژگی هدف را در مجموعه داده‌ها مشخص کنیم. این الگوریتم کار خود را از گره ریشه آغاز می‌کند. داده‌های آموزشی به دو گره فرزند و سپس هر گره فرزند به ترتیب به دو گره نوه تقسیم می‌شوند. در فرایند رشد درخت، الگوریتم CART در جستجوی سؤالاتی است که هر گره والد را به دو گره فرزند همگن تقسیم نماید (۲۶). منظور از همگن بودن گره این است که همه رکورد‌های موجود در آن متعلق به یک دسته خاص باشند؛ چون در این صورت آن گره به برگ تبدیل می‌شود. به‌عبارت دیگر الگوریتم مذکور به دنبال ویژگی‌هایی از مجموعه داده‌ها است که خاصیت جدا کنندگی بیشتری دارند (۲۶). هر چه درخت بیشتر رشد کند، گره‌ها همگن‌تر می‌شوند و اطلاعات بیشتر نمایان می‌گردد (۲۶). فرایند رشد درخت تا هنگام رسیدن به درختی با اندازه ماکزیمم و تا زمانی که عملیات تقسیم به دلیل کمبود داده‌ها متوقف نشود، ادامه خواهد داشت (۱۴).

پس از ساخت درخت، عملیات هرس کردن درخت با اندازه ماکزیمم توسط یکی از روش‌های هرس کردن با شروع از

مرحله آموزش، الگوریتم یادگیرنده براساس مجموعه داده‌های آموزشی یک مدل می‌سازد. شکل مدل ساخته شده به الگوریتم یادگیرنده مورد استفاده بستگی دارد که در این پژوهش از الگوریتم درخت تصمیم CART استفاده شده است، بنابراین مدل ساخته شده، یک درخت تصمیم خواهد بود. در مرحله ارزیابی (آزمایشی)، براساس مجموعه داده‌های آزمایشی دقت و کارایی مدل ساخته شده ارزیابی خواهد شد. داده‌هایی که در مرحله ارزیابی مورد استفاده قرار می‌گیرند در مرحله آموزش و برای ساخت مدل استفاده نشده‌اند. در واقع، این داده‌ها برای مدل ساخته شده، پیش از این ناشناخته هستند. یعنی با وجود این که الگوریتم دسته‌بندی برچسب رکورد‌های آزمایشی را در اختیار دارد (جهت محاسبه دقت مدل ساخته شده پس از فرایند یادگیری مدل)، ولی اجازه استفاده از این رکورد‌ها را در مرحله آموزش ندارد. این روال در شکل زیر نشان داده شده است:

نمایش دانش به شکل درخت سبب شده است که دسته‌بندی‌های مبتنی بر درخت تصمیم کاملاً قابل تفسیر باشند (۲۶). درخت تصمیم، ساختار درختی شبیه فلوچارت دارد به گونه‌ای که هر گره داخلی نمایانگر یک آزمون بر روی یک ویژگی، هر شاخه به معنای خروجی آزمون است و برچسب‌های کلاس توسط برگ‌ها نشان داده می‌شوند. درخت‌های تصمیم در حوزه‌های مختلفی از زندگی روزمره از جمله تجارت، مدلسازی انرژی، پزشکی، پردازش تصویر، صنعت و ... کاربرد دارند (۲۶).



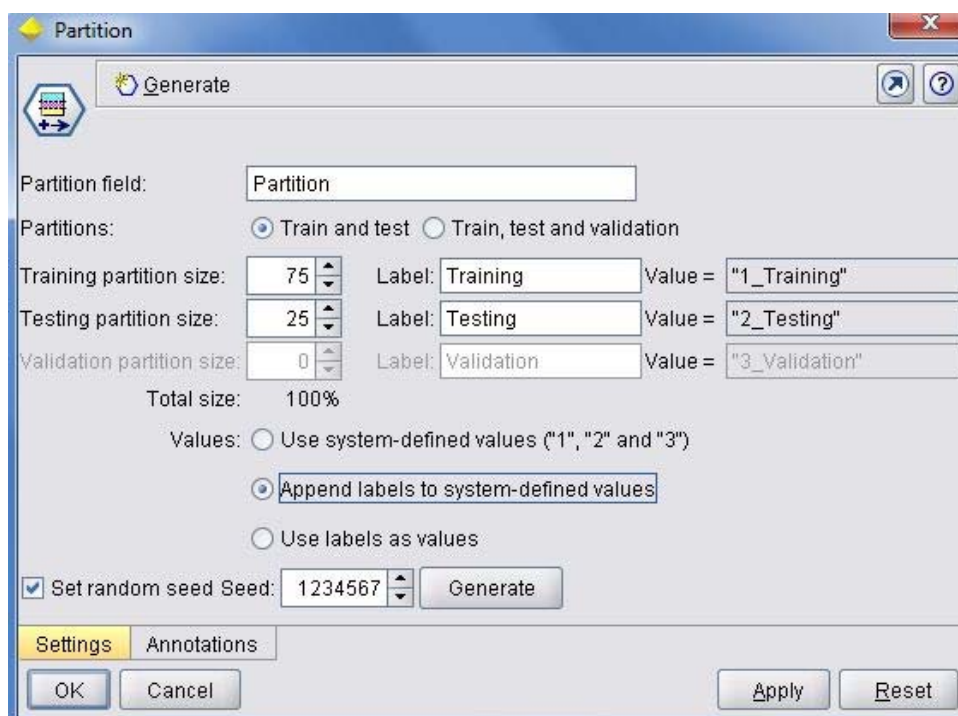
شکل ۲- مراحل فرایند دسته‌بندی

یافته‌ها

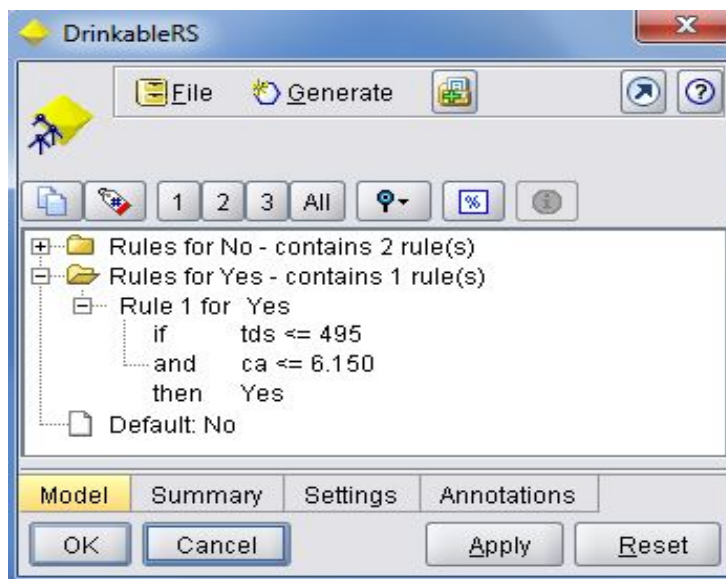
همان‌طور که در بخش دسته‌بندی بیان شد، پیش از ورود داده‌ها به الگوریتم درخت تصمیم، تفکیک داده‌ها به مجموعه داده‌های آموزشی و آزمایشی انجام شد. در این مرحله و به‌طور تصادفی ۷۵ درصد از نمونه‌ها در مجموعه داده‌های آموزشی و ۲۵ درصد از نمونه‌ها در مجموعه داده‌های آزمایشی قرار گرفتند. در پایان این مرحله که با استفاده از نرم‌افزار Clementine 12.0 انجام شد، تعداد ۴۳ نمونه در مجموعه داده‌های آموزشی و تعداد ۱۷ نمونه در مجموعه داده‌های آزمایشی قرار گرفتند (شکل ۳).

پس از تفکیک داده‌ها به مجموعه داده‌های آموزشی و آزمایشی، اجرای مدل براساس الگوریتم درخت تصمیم CART انجام شد که مؤثرترین قاعده آن در شکل ۴ نشان داده شده است. از این قانون می‌توان به منظور تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی استفاده نمود. تفسیر این قانون، به شرح زیر است:

برگ‌ها به سمت ریشه انجام می‌شود. مکانیسم CART قصد تولید تنها یک درخت را ندارد، بلکه تلاش می‌کند تا یک توالی از درخت‌های هرس شده تو در تو را ایجاد نماید به طوری که هر یک از آنها کاندیدایی هستند که می‌توانند در نهایت به عنوان درخت بهینه انتخاب شوند. یک درخت خوب به‌وسیله ارزیابی کارایی آن بر روی داده‌های آزمایشی مستقل شناسایی می‌گردد. در نهایت می‌توان درخت تصمیم ساخته شده را بر روی مجموعه داده‌های آزمایشی اعمال نمود. منظور از اعمال کردن مدل، پیش‌بینی مقدار ویژگی دسته برای یک رکورد آزمایشی براساس مدل ساخته شده است (۲۶). شایان ذکر است این روند ارزیابی که شامل مقایسه هر یک از مقادیر موجود در ویژگی‌های هر رکورد آزمایشی با قواعد موجود در مدل ساخته شده است نیز در نرم‌افزار Clementine 12.0 موجود است تا این عملیات زمان‌بر به صورت خودکار و به سرعت انجام پذیرد.



شکل ۳- تقسیم مجموعه داده‌ها به دو مجموعه داده‌های آموزشی و آزمایشی



شکل ۴- قانون استخراج شده جهت تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی

پس از ساخت مدل و استخراج قواعد فوق (شکل ۴)، با استفاده از الگوریتم درخت تصمیم CART، نتایج به دست آمده مورد ارزیابی قرار گرفت تا میزان اطمینان و صحت آنها تعیین گردند. برای این منظور، آن دسته از نمونه‌هایی را که در مجموعه داده‌های آزمایشی قرار گرفتند (که الگوریتم برای ساخت مدل از آنها استفاده نکرده است)، به عنوان ورودی به مدل ساخته

مؤثرترین فاکتورهای تاثیرگذار در کیفیت آب آشامیدنی در دشت کازرون (شکل ۴) عبارتند از: کل جامدات محلول (TDS) و کلسیم (Ca). بدین ترتیب، در صورتی که TDS در این دشت کمتر یا مساوی ۴۹۵ L/mg و Ca آن، کمتر یا مساوی ۶/۱۵۰ L/meq باشد، این آب برای آشامیدن مناسب است.

Results for output field Drinkable			
Comparing \$R-Drinkable with Drinkable			
'Partition'	2_Testing		
Correct	16	94.12%	
Wrong	1	5.88%	
Total	17		

شکل ۵- نتیجه اولیه ارزیابی انجام شده از الگوریتم CART

Results for output field Drinkable		
Comparing \$R-Drinkable with Drinkable		
Correct	54	90%
Wrong	6	10%
Total	60	

شکل ۶- نتیجه نهایی ارزیابی انجام شده از الگوریتم CART

بحث

با تحلیل مجموعه داده‌ها و پارامترهای انتخاب شده و با بهره‌گیری از تکنیک داده کاوی درخت تصمیم و الگوریتم CART مؤثرترین فاکتورهای کیفیت آب آشامیدنی در دشت کازرون شناسایی گردیدند. با توجه به دو عامل کل جامدات محلول (TDS) و کلسیم (Ca)، به نظر می‌رسد این دو فاکتور تاثیر بیشتری بر کیفیت آب آشامیدنی در این دشت دارند. دلیل اثرگذاری بیشتر این دو عامل را می‌توان به ساختار سازندهای زمین‌شناسی منطقه و وجود کربنات کلسیم (آهک)، در ترکیب آنها مرتبط دانست. این نتایج مانند نتایج پژوهش‌های Fu-Cheng و همکار (۱۹)، Azhar و همکاران (۲۰)، Gu و همکاران (۲۱)، Lee و همکار (۲۲)، Thompson (۲۳)، کاربرد تکنیک‌های داده کاوی در تعیین مؤثرترین فاکتورهای کیفیت آب را تایید می‌کند. همچنین نتایج این پژوهش با یافته‌های Arora و همکاران (۱۶)، Ji و همکاران (۱۸)، و Soleimanpour و همکاران (۵، ۱۰)، مبنی بر کارایی روش

شده وارد شد تا مشاهده شود که مدل ساخته شده تا چه حد قادر به تشخیص برچسب نمونه‌های آزمایشی خواهد بود. همان‌طور که در شکل ۵ نشان داده شده است، نتیجه ارزیابی مدل ساخته شده دارای اطمینان ۹۴/۱۲ درصد است. به عبارت دیگر، مدل ساخته شده برای ۱۶ نمونه آزمایشی قادر به تشخیص صحیح برچسب آنها شده است و تنها برای یک نمونه آزمایشی ناموفق عمل کرده است.

به‌منظور اطمینان بیشتر، کل مجموعه داده‌ها (آموزشی و آزمایشی) به صورت یک جا به مدل ساخته شده وارد شد تا ارزیابی دقیق‌تری صورت پذیرد. در این مرحله، مدل ساخته شده براساس الگوریتم درخت تصمیم CART موفق به تشخیص صحیح ۵۴ نمونه از ۶۰ نمونه وارد شده گردید. بنابراین همان‌طور که در شکل ۶ نشان داده شده است، دقت به‌دست آمده براساس مؤثرترین فاکتورهای عدم کیفیت آب آشامیدنی در مدل نهایی، برابر با ۹۰ درصد است.

خارج از باند معرفی شده نماید؛ اما در روش‌های آماری چنین امکانی وجود نداشته و ترم‌های خطا به سختی قابل تغییر و سفارشی‌سازی هستند. بنابراین مدل‌ها بیش از اندازه کلی و غیر حساس به تغییرات هستند.

با عنایت به مباحث فوق، می‌توان بیان داشت امروزه استفاده مؤثر و استخراج داده‌های پنهان از انبوه داده‌ها توسط مدیران به‌عنوان یک هدف عمده جهت بهبود وضعیت موجود تلقی می‌شود؛ و استفاده از تکنیک‌های داده‌کاوی می‌تواند به عنوان ابزار عملیاتی در شناسایی مناطق آلوده و ارزیابی و نظارت بر کیفیت منابع آب‌های سطحی و زیرزمینی، و مطالعه و توسعه منابع آب، به منظور تصمیم‌گیری‌های علمی برنامه‌ریزان و مدیران مفید واقع شود.

نتیجه‌گیری

با عنایت به مباحث فوق، می‌توان بیان داشت امروزه استفاده مؤثر و استخراج داده‌های پنهان از انبوه داده‌ها توسط مدیران به‌عنوان یک هدف عمده جهت بهبود وضعیت موجود تلقی می‌شود؛ و استفاده از تکنیک‌های داده‌کاوی می‌تواند به عنوان ابزار عملیاتی در شناسایی مناطق آلوده و ارزیابی و نظارت بر کیفیت منابع آب‌های سطحی و زیرزمینی، و مطالعه و توسعه منابع آب، به منظور تصمیم‌گیری‌های علمی برنامه‌ریزان و مدیران مفید واقع شود.

توصیه می‌گردد با توجه به کاربرد مفید دانش داده‌کاوی در مباحث منابع آب، نسبت به آماربرداری‌های دقیق از این منابع و ایجاد یک بانک اطلاعاتی جامع از آمار مستند از منابع آب کشور اقدام عملی انجام شود تا بتوان با در اختیار داشتن داده‌هایی صحیح و مطمئن، با اطمینان بیشتری از این دانش جدید بهره جست. در پایان با عنایت به نتایج این پژوهش پیشنهاد می‌گردد نسبت به تصفیه و کاهش املاح محلول موجود در آب توجه جدی شود؛ همچنین بر انجام پایش‌های مستمر در قالب نمونه‌برداری‌های دوره‌ای منظم از چاه‌های این دشت تاکید می‌گردد.

CART کاملاً همخوانی دارد؛ همچنین این نتایج همانند مطالعات Cho (۱۷)، و Rajakumari و همکار (۱۱)، مؤید کاربرد تکنیک‌های داده‌کاوی، و تاییدکننده اثربخشی این روش در تحقیقات مرتبط با منابع آب و به‌ویژه در حوزه کنترل کیفی آب است.

همچنین توجه به این نکته ضروری است که با توجه به این‌که تقریباً در اغلب مطالعات حوزه کیفی منابع آب با استفاده از روش‌های آماری به بررسی رابطه بین پارامترها و بررسی همبستگی بین آنها پرداخته می‌شود؛ اما در عالم واقعیت ممکن است چندین پدیده به‌صورت همزمان بر پارامتری اثر کنند و الگویی را به وجود بیاورند؛ بنابراین بررسی رابطه دو دویی پارامترها، در این مطالعات گاهی بسیار دور از واقعیت و ابتدایی است؛ این در حالی است که با افزایش تعداد پارامترها، روش‌های آماری، توانایی یافتن الگوها را از دست می‌دهند و به‌علت ماهیت اغلب خطی خود، از کشف روابط غیر خطی و پیچیده بین متغیرها عاجز هستند؛ اما روش‌های داده‌کاوی به نحوی طراحی شده‌اند که می‌توانند روابط مرکب و پیچیده بین چندین پارامتر را در پایگاه داده کشف کنند. بنابراین روش‌های آماری، توانایی کشف الگوهای پیچیده و غیر خطی را ندارند؛ در حالی که روش‌های داده‌کاوی به علت خاصیت اکتشافی، بدون هیچ فرض اولیه‌ای شروع به مدلسازی رفتار داده‌ها می‌نمایند و به مرور زمان و با جلو رفتن الگوریتم، الگو پررنگ‌تر و پررنگ‌تر خواهد گردید. ساختار غیر خطی و مقاوم این مدل‌ها، توانایی شبیه‌سازی رفتار محیط‌های واقعی (نمونه عینی آن عرصه‌ها و محیط‌های طبیعی) را به روش‌های داده‌کاوی می‌دهد. همچنین مزیت برتر استفاده از روش‌های داده‌کاوی نسبت به روش‌های آماری در مطالعات منابع آب آن است که روش‌های داده‌کاوی معمولاً نسبت به تنظیم پارامترها حساسیت کمتری نسبت به روش‌های آماری دارند و دارای ساختار انعطاف‌پذیرتری هستند. به‌علاوه در تکنیک‌های داده‌کاوی می‌توان الگوریتم‌ها را به نحوی تنظیم کرد که "خطا از حدی کمتر" را جزء خطا به حساب نیاید و سامانه تمام تلاش خود را صرف کاهش خطاهای بزرگ و

ملاحظات اخلاقی

نویسندگان کلیه نکات اخلاقی شامل عدم سرقت ادبی، انتشار دوگانه، تحریف داده‌ها و داده‌سازی را در این مقاله رعایت کرده‌اند.

References

1. Bower AS, Serra N, Ambar I. Structure of the Mediterranean Undercurrent and Mediterranean Water spreading around the southwestern Iberian Peninsula. *Journal of Geophysical Research: Oceans*. 2002;107(C10):1-19.
2. Shareateshirenasab A, Soleimanpour S, Jowkar L. Relationship between water quality factors using stepwise in the Zareendasht region, Fars province. *Proceedings of the First National Congress of Desert*; 2012; Karaj, Iran (in Persian).
3. Soleimanpour S, Shareateshirenasab A. Investigation of the quality of groundwater resources and the changes of qualitative elements (Case Study: Plain Khosouyeh Sachun plain, Zareendasht region, Fars province). *Proceedings of the First National Congress of Desert*; 2012; Karaj, Iran (in Persian).
4. FAO. *State of the World's Forests 2007*. Rome: Food and Agriculture Organization of the United Nations; 2007.
5. Soleimanpour S, Hedayati B, Zolfaghari M. Determination of effective factors of drinking water quality by using the QUEST data mining technique in Saadatsharh- Fars Province. *Proceedings of 3rd International Conference on Rainwater Catchment Systems*; 2015; Birjand, Iran (in Persian).
6. Ahuja S. Regionalization of river basins using cluster ensemble. *Journal of Water Resource and Protection*. 2012;4(07):560-66.
7. Salajegheh A, Razavezadeh S, Khorasani N, Hamidifar M, Salajegheh S. Land use changes and its effects on water quality (Case study: Karkheh watershed). *Journal of Environmental Studies* 37(58):81-86 (in Persian).
8. Sepehrnia B, Nabizadeh R, Mahvi A, Naseri S. Water quality analysis of drinking water distribution systems of Rey Township using IWQIS software. *Iranian Journal of Health & Environment*. 2016;9(1):103-14.
9. Zare Abyaneh H. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*. 2014;12(1):40.
10. Soleimanpour S, Mesbah S, Hedayati B. The application of data mining algorithm K-Means and CART most influential factors in determining the quality of drinking water in Nurabad plain of Fars Province. *Proceedings of Eleventh National Congress of Watershed Management Science and Engineering*; 2016; Yasouj, Iran (in Persian).
11. Rajakumari SB, Nalini C. Identification of lead contaminant in river water quality data. *Journal of Chemical and Pharmaceutical Sciences*. 2016;9(4):2764-66.
12. Navi M. *Identify load components with the use of data mining techniques [dissertation]*. Tehran: Tarbiat Modares University; 2007 (in Persian).
13. Shoba G, Shobha G. Water Quality prediction using data mining techniques: A survey. *International Journal of Engineering and Computer Science*. 2014;3(6):6299-306.
14. Shahrabi J. *Data Mining*. Tehran: Jahad-e Daneshgahi Publication; 2013 (in Persian).
15. Cichosz P. *Data Mining Algorithms: Explained Using R*. New York: John Wiley & Sons; 2014.
16. Arora N, Arora AS, Sharma S, Reddy AS. Use of cluster analysis-A data mining tool for improved water quality monitoring of River Satluj. *International Journal of Advanced Computer Science and Applications*. 2014;6(2):63-69.
17. Cho Y. A watershed water quality evaluation model using data mining as an alternative to physical watershed models. *Water Science and Technology: Water Supply*. 2016;16(3):703-14.

18. Ji Y, Lei X, Cai S, Wang X. Application of a classifier based on data mining techniques in water supply operation. *Water*. 2016;8(12):599.
19. Fu-Cheng L, Xue-Zhao H. Application of Fuzzy c-means Clustering for assessing rural surface water quality in Lianyungang City. 2013 Proceedings of Fifth International Conference on Measuring Technology and Mechatronics Automation; 2013; Hong Kong. New York: Institute of Electrical and Electronics Engineers; 2013.
20. Azhar SAS, Johar H, Baki SRMS, Tahir NM. Optimization of water quality monitoring based on fuzzy algorithms. Proceedings of IEEE Conference on Systems, Process & Control (ICSPC); 2013; Malaysia. New York: Institute of Electrical and Electronics Engineers; 2013.
21. Gu Q, Deng J, Wang K, Lin Y, Li J, Gan M, et al. Identification and assessment of potential water quality impact factors for drinking-water reservoirs. *International Journal of Environmental Research and Public Health*. 2014;11(6):6069-84.
22. Lee S, Lee C-W. Application of decision-tree model to groundwater productivity-potential mapping. *Sustainability*. 2015;7(10):13416-32.
23. Thompson E. Investigating drinking water advisories in first nations communities through data mining [dissertation]. Canada: University of Guelph; 2016.
24. Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining. Boston: Pearson Addison Wesley; 2005.
25. Sanieabadeh M, Mahmoodi S, Taherpour M. Applied Data Mining. Tehran: Niازه Danesh; 2012 (in Persian).
26. Sharma H, Kumar S. A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research*. 2016;5(4):2094-97.
27. Gordon L. Using classification and regression trees (CART) in SAS® enterprise miner TM for applications in public health. *Public Health*. 2013.
28. Loh WY. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2011;1(1):14-23.



Application of CART decision tree data mining to determine the most effective drinking water quality factors (case study: Kazeroon plain, Fars province)

SM Soleimanpour^{1,*}, SH Mesbah¹, B Hedayati²

1- Soil Conservation and Watershed Management Research Department, Fars Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Shiraz, Iran

2- Computer Engineering, Payame Noor University, Tehran, Iran

ARTICLE INFORMATION:

Received: 28 November 2017

Revised: 17 February 2018

Accepted: 21 February 2018

Published: 20 June 2018

Keywords: Drinking water, Data mining, Decision tree, Kazeroon plain, Water quality

***Corresponding Author:**
m.soleimanpour@yahoo.com

ABSTRACT

Background and Objective: Determination of quality parameters of drinkable water is important, especially in developing countries, to increase the productivity and better management and planning of water resources. The aim of current study was to apply CART decision tree data mining technique to determine the most effective factors on drinkable water quality in Kazeroon plain, located west of Fars province, Iran.

Materials and Methods: Qualitative parameters of 60 drinkable wells such as SAR, Na, Cl, SO₄, TH, TDS, pH, NO₃, CaCO₃, HCO₃, Ca, Mg, K and EC were taken in the study area. The most effective factors on quality of drinkable water were determined with 90% accuracy, using CART decision tree data mining technique in Clementine 12.0 software.

Results: The results showed that total dissolved solids (TDS) and calcium content (Ca) had the highest impact on quality of drinking water. Therefore, when the TDS of water in this plain is equal or less than 495 mg/L and the calcium content is equal or less than 6.150 meq/L, the water is suitable for drinking.

Conclusion: The TDS and Ca content were the most effective parameters on the quality of drinkable water in this plain, due to its geological formation and the existence of CaCO₃ in its structure. The water purification, reduction of soluble material concentration, and monitoring of wells in this plain are recommended.

Please cite this article as: Soleimanpour SM, Mesbah SH, Hedayati B. Application of CART decision tree data mining to determine the most effective drinking water quality factors (case study: Kazeroon plain, Fars province). Iranian Journal of Health and Environment. 2018;11(1):1-14.