مجله سلامت و محیط زیست، فصلنامه‌ی علمی پژوهشی
انجمن علمی بهداشت محیط ایران
دوره یازدهم، شماره اول، بهار ۱۳۹۷، صفحات ۱ تا ۱۴

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[1*]، سید حمید مصباح[1]، بهرام هدایتی[2]
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:

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

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 موثر 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، Na، Cl، SO_4، TH، TDS، pH، NO_3، $CaCO_3$، HCO_3، 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.

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.

مقدمه

در کشـورهای غربی و صنعتی سـالانه هر فرد حداقل میزان به ۲۰۰۰ m^3 آب برای برخورداری از یک اسـتاندارد مطلوب نیاز دارد. اگر سـرانه آب هر فرد بین ۱۰۰۰ تا ۲۰۰۰ m^3 باشد آن کشور تحت تنش آب است. اما در صورتی که سرانه از ۵۰۰ m^3 در سـال کمتر باشد، کشـور مذکور با کم آبی مواجه است. در حال حاضر منابع آب موجود می‌تواند سـالانه ۷۰۰۰ m^3 برای هر فرد آب فراهم نماید. اگرچه آب کافی حداقل برای سه برابر جمعیت کره زمین موجود اسـت، ولی عـدم تعادل بین توزیع جمعیـت و بارندگـی، موجب کمبود آب و اسـتفاده از آب‌های آلوده در بعضی از مناطق شـده اسـت (۱)؛ از این‌رو، وجود آب آشـامیدنی سالم، ضامن سـلامتی جامعه است و اولین قدم در دسـتیابی به این مهم، بررسـی پارامترهای آن است (۲). علاوه بر این، مهمترین مسائلی که در اثر مصرف آب‌های نامناسب در امور کشـاورزی ایجاد می‌شود شامل شور شدن ثانویه خاک‌ها، کاهش نفوذپذیری و سـمیت املاح است؛ که هر کدام به نوعی بر سلامتی و یا رشـد و تولید محصولات (کشاورزی، دامدارای، صنعتی و ...) و در نهایت آسـیب به سیسـتم ایمنی و سـلامت انسان خدشه وارد می‌نماید (۳).

وضعیت منابع آب شـرب (به‌ویژه منابع شرب سالم)، در جهان روز به روز در حال کاهش اسـت. سـازمان خواربار و کشاورزی ملـل متحد، علاوه بر اعلام نگرانـی در این خصوص، پیش‌بینی کرده دو سـوم جمعیت جهان در سال ۲۰۲۵ میلادی با کمبود آب روبرو خواهند بود (۴).

ایران، کشـوری پهناور است که از نظر منابع آبی محدود است. بـه همیـن جهت از زمان‌های دور، آب در کشـور ما از ارزش و اهمیت والایی برخوردار بوده است. شرایط اقلیمی، زمین‌شناسی و آب‌شناسـی کشور ما نیز طوری است که بهره‌برداری از آب با کیفیت را با مسائل و مشکلات خاص خود همراه می‌کند. انسان برای آشامیدن، بهداشـت فردی و اجتماعی، پخت و پز، شست و شـو و آبیاری به آب مناسـب و با کیفیـت احتیاج دارد و هر انسـان در روز حداقل به ۲/۵ L آب شـیرین سالم و با کیفیت اسـتاندارد برای آشـامیدن نیاز دارد تا بتواند احتیاجات خود را برطرف نماید. علاوه بر آن، برای هر نفر در کشـور ایران به‌طور متوسـط روزانه به ۱۱۰ L آب برای مصارف شـخصی و خانگی نیاز است که در واقع همان سرانه مصرف آب در شهرهای کشور اسـت (۵). به این منظور و در پی کمبود جهانی آب، پیش‌بینی کمیت و کیفیت آب، توجه جدی پژوهشگران هیدرولوژی را به خود جلب کرده اسـت (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 بـه تنهایـی گویای وضعیت فعلی آب باشـد و برای تشـخیص کیفیت آب بایسـتی توجه بیشـتری به شاخص‌های کیفی آب معطـوف نمود و کاربرد این شـاخص‌ها را به صورت یک معیار ضروری جهت ارزیابی کیفیت آب شـرب در شـبکه‌های توزیع و منابـع تامین آب مورد توجه قرار داد (۸). در نتیجه بررسـی پارامترهای کیفیـت آب به منظور افزایش بهره‌وری و توسـعه مدیریت و برنامه‌ریزی بهتر منابع، امری ضروری اسـت (۹)؛ به این منظور یکی از راه‌هـای اثربخش و تاثیرگذار برای مبارزه با این معضل ملی، اسـتفاده بهینه از منابع آبی کشـور با تاکید بر اصول کیفیت آب اسـت، به‌طوری که امروزه بررسی‌های کیفی آب، دامنه گسـترده‌ای پیدا کرده و مسـائل مربوط به آلودگی آب‌های سـطحی 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 اسـت (۱۱). این علـم، به‌عنوان مجموعه‌ای از روش‌ها در فرایند کشف دانش است که برای تشخیص الگوها و رابطه‌های نامعلوم در داده‌ها مورد اسـتفاده قرار می‌گیرد (۱۲). به‌عبارت دیگر، داده‌کاوی مجموعه‌ای از فعالیت‌هایی اسـت که بـرای یافتـن الگوهای جدید، پنهان، و غیـر منتظره در داده‌ها استفاده می‌شود (۱۳).

از دلائل رشـد و توسـعه دانـش داده‌کاوی در علـوم مختلف، می‌تـوان به افزایـش حجم پایگاه داده‌ها و عدم توانایی انسـان برای درک و اسـتخراج از آنها اشـاره کـرد. بنابراین درک این

داده‌ها بدون ابزارهای قدرتمند میسر نیست؛ زیرا تصمیم‌گیران ابزار قوی برای استخراج اطلاعات با ارزش را در دست ندارند. در واقع شرایط فعلی توصیف کننده 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. شیوه نمایش 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 زمانی‌که 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 صورتی‌که سختی کل در این دشت بین ۱۷۵ و ۴۴۰ ppm، و هدایت الکتریکی عصاره اشباع آن، کمتر یا مساوی ۷۳۱/۶۵ µmhos/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. به علاوه 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 ML (Multi Layer) 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.

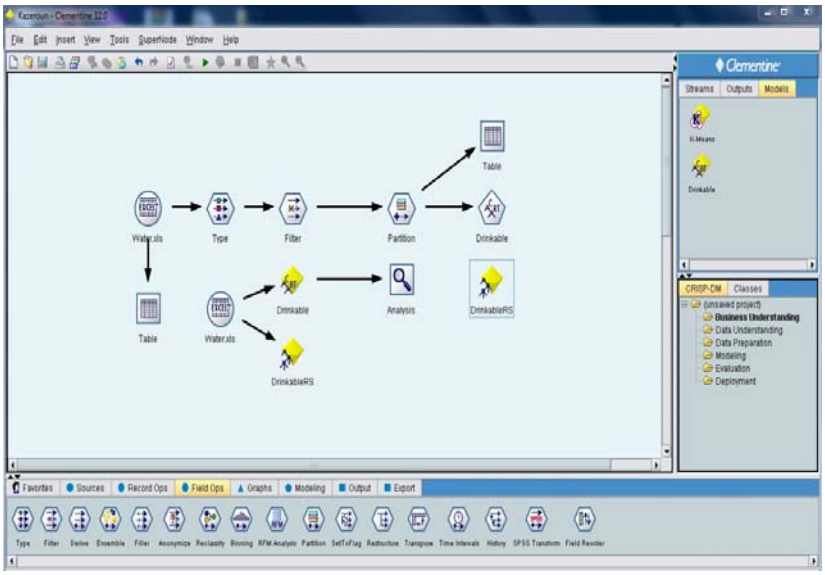
Fu-Cheng و همکار (۱۹) از روش خوشه‌بندی فازی

c-means بــه منظــور طبقه‌بندی و ارزیابــی کیفیت آب‌های ســطحی روستایی ۳۳ ایستگاه در شــهر لیان یانگانگ استفاده و مناطــق آلوده به آمونیاک و نیتروژن را تعیین کردند. ایشــان کاربرد این روش را در شناســایی مناطق آلوده و ارزیابی کیفی منابــع آب، مطلوب ارزیابی نمودنــد. Azhar و همکاران (۲۰) هم از منطق فازی و درخت تصمیم به صورت توأم در بررســی استانداردهای کیفیت آب استفاده کردند و اعلام داشتند به‌منظور نظارت بر کیفیت آب، بهره‌گیری از این روش‌ها مناســب است. به‌علاوه Gu و همکاران (۲۱) در پژوهشی به شناسایی و ارزیابی عوامــل تاثیرگــذار بر کیفیت آب در ۷۳ مخزن آب آشــامیدنی در اســتان ژجیانگ چین اقدام کردنــد. نتایج این تحقیق که با بهره‌گیری از درخت تصمیم‌گیری CART، به‌دست آمد بیانگر ۸۱ درصــد دقت در تخمین کیفیت آب بــود. با توجه به نتایج مطلوب، ایشــان اســتفاده از این روش را در مطالعه کیفیت آب آشامیدنی ســایر مناطق پیشــنهاد و بیان کردند که این روش می‌تواند به‌طور بالقوه به‌عنوان ابزار عملیاتی برای برنامه‌ریزان و مدیــران عمل نماید. همچنین Lee و همکار (۲۲) در تحقیقی پتانسیل بهره‌وری آب زیرزمینی را با استفاده از تکنیک درخت تصمیم و سیســتم اطلاعات جغرافیایی در شــهرهای بورایونگ و پوهانــگ کره، مورد بررســی قرار دادند. نتایج نشــان داد که مدل‌های درخت تصمیم‌گیری می‌توانند برای مطالعه و توســعه منابــع آب زیرزمینی مفید باشــند. Thompson (۲۳) نیز در پژوهشــی با اســتفاده از تکنیک‌های درخت تصمیم به بررسی کیفیــت آب آشــامیدنی در کانادا پرداخت و اعــلام نمود نتایج حاصــل از روش‌هــای داده‌کاوی و درخت تصمیــم 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 اســت. زیرا این امر ارتباط بســیاری با سلامتی انسان و جامعه بشری دارد. به طوری‌که بررسی کیفیت آب (به ویژه آب آشامیدنی)، و چگونگی تغییر آنها نقش بسیار مهمی در مدیریت و بهره‌بــرداری از منابــع دارد. به‌علاوه با آگاهــی از فاکتورهای مؤثر در تغییر کیفیت آب آشــامیدنی می‌توان در جهت هر چه بهتــر مدیریت کردن آنها اقدامات مؤثری را به عمل آورد. بدین منظور پژوهش حاضر نسبت به کاربرد تکنیک داده‌کاوی درخت تصمیــم CART در تعیین مؤثرتریــن فاکتورهای کیفیت آب آشــامیدنی در دشــت کازرون واقع در غرب استان فارس اقدام نموده است.

مواد و روش‌ها

به‌منظور انجام این پژوهش، با مراجعه به سازمان آب منطقه‌ای استان فارس، نســبت 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 "۵۱°۳۹'۱۵ طول شرقی، و '۲۹°۲۷ عرض شــمالی واقع شده است. آمارهای اخذ شده برای ۶۰ حلقه چاه، مشــتمل بر فاکتورهای کیفــی SAR، Na، Cl، SO_4، TH، TDS، pH، NO_3، $CaCO_3$، HCO_3، Ca، Mg، K و EC بود. به‌منظور نرمال بودن داده‌ها، از نرم افزار SPSS16، و آزمون کولموگــروف- اســمیرنوف (Kolmogorov-Smirnov) استفاده شــد و نرمال بودن داده‌ها تایید شــد. سپس به‌منظور تعییــن مؤثرترین فاکتورهای کیفیت آب آشــامیدنی از تکنیک داده‌کاوی درخت تصمیم CART اســتفاده شد. مدل‌سازی در نرم افزار Clementine 12.0 انجام گرفت. نرم افزار یاد شده ســاخت شرکت SPSS است و امکان ایجاد مدل‌های متعددی را براســاس تئوری‌هــای آمــاری، هوش مصنوعــی 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، کلیــه 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 باشــد تا در نهایت هدفی را برآورده ســازند. هر گره در جریان، وظیفه‌ای بر عهده دارد و مانند یک



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

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 این‌که یک مجموعه داده (Dataset)، ممکن است شامل اطلاعات زیادی باشد و اصولا 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

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

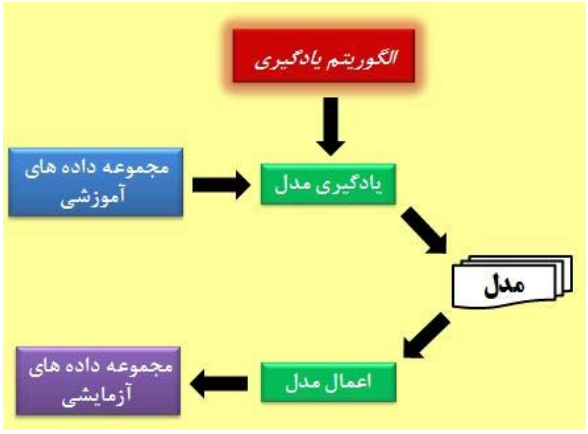
در الگوریتم‌های دسته‌بندی، کل مجموعه داده‌ها به دو قسمت مجموعه داده‌های آموزشی و مجموعه داده‌های آزمایشی تقسیم می‌شوند (۲۴). همان‌طور که در شکل ۲ نشان داده شده است، الگوریتم‌های دسته‌بندی شامل دو مرحله‌ی آموزش و آزمایش هستند. در مرحله آموزش، الگوریتم یادگیرنده براساس مجموعه داده‌های آموزشی، یک مدل را 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 مدل 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

جدول ۱- فاکتورهای کیفی آب آشامیدنی

واحد	فاکتورهای کیفی و محدوده‌های مجاز
meq/L	۰< $CaCO_3$ <۵۰۰
meq/L	۰< HCO_3 <۵۰۰
meq/L	۰< NO_3 <۵۰
meq/L	۰< SO_4 <۴۰۰
meq/L	۰< Ca <۳۰۰
meq/L	۰< Cl <۴۰۰
meq/L	۰< K <۱۲
meq/L	۰< Mg <۱۵۰
meq/L	۰< Na <۲۰۰
µmhos/cm	۰< EC <۷۵۰
mg/L	۰< TDS <۱۵۰۰
ppm	۰< TH <۵۰۰
-	۶/۵< pH <۹
-	۰< SAR <۱۸

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 کاملا 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

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 زمانی‌که 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

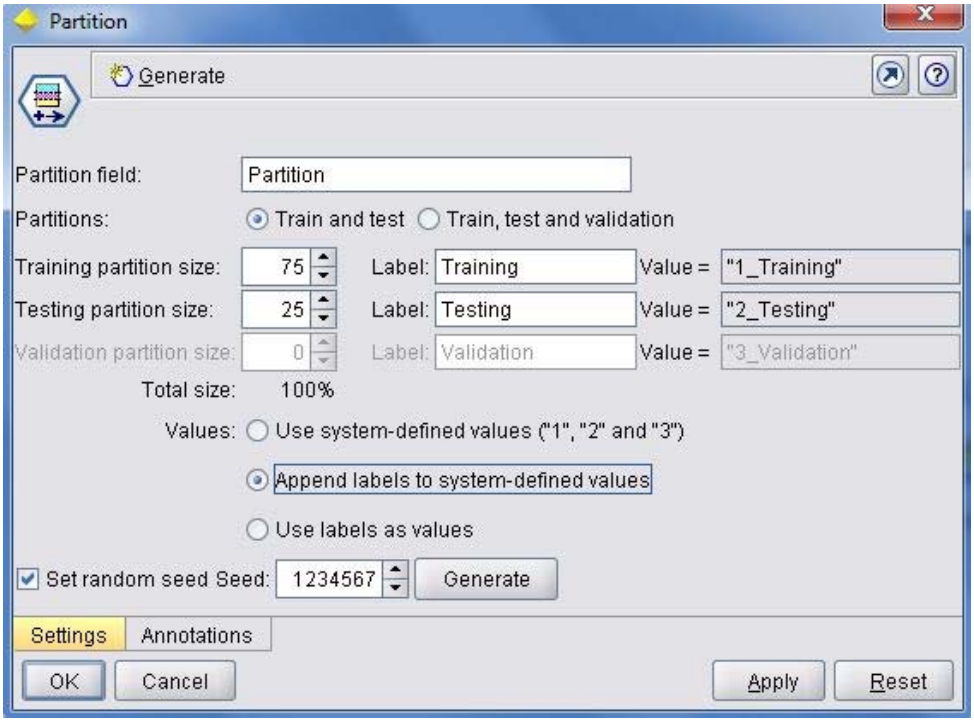
برگ‌ها به سمت ریشــه انجام می‌شود. مکانیسم CART قصد تولید تنها یک درخت را ندارد، بلکه تلاش می‌کند تا یک توالی از درخت‌های هرس شــده تو در تــو را ایجاد نماید به طوری‌که هــر یک از آنها کاندیدایی هســتند که می‌تواننــد در نهایت به عنوان درخت بهینه انتخاب شــوند. یک درخت خوب به‌وسیله ارزیابی کارایی آن بر روی داده‌های آزمایشی مستقل شناسایی می‌گردد. در نهایت می‌توان درخت تصمیم ســاخته شــده را بر روی مجموعه داده‌های آزمایشــی اعمال نمود. منظور از اعمال کــردن مدل، پیش‌بینی مقدار ویژگی دســته بــرای یک رکورد آزمایشــی براســاس مدل ساخته شده اســت (۲۶). شایان ذکر اســت این روند ارزیابی که شــامل مقایســه هر یــک از مقادیر موجــود در ویژگی‌های هر رکورد آزمایشــی با قواعد موجود در مدل ســاخته شده است نیز در نرم افزار Clementine 12.0 موجود اســت تا این عملیــات زمان‌بر به صورت خــودکار و به سرعت انجام پذیرد.

یافته‌ها

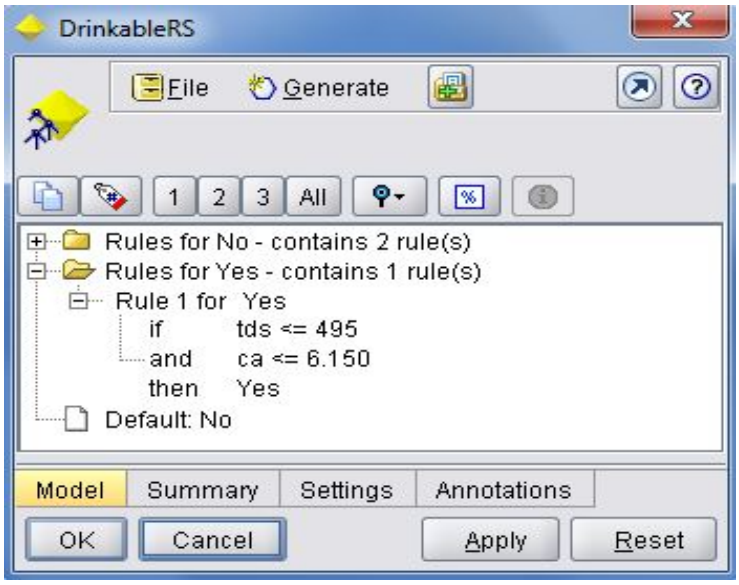
همان‌طور که در بخش دسته‌بندی بیان شد، پیش از ورود داده‌ها به الگوریتم درخت تصمیم، تفکیک داده‌ها به مجموعه داده‌های آموزشــی و آزمایشی انجام شد. در این مرحله و به‌طور تصادفی ۷۵ درصد از نمونه‌ها در مجموعه داده‌های آموزشی و ۲۵ درصد از نمونه‌ها در مجموعه داده‌های آزمایشی قرار گرفتند. در پایان این مرحله که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

پس از تفکیک داده‌ها به مجموعه داده‌های آموزشی و آزمایشی، اجرای مدل براســاس الگوریتم درخــت تصمیم CART انجام شد که مؤثرترین قاعده آن در شکل ۴ نشان داده شده است.

از ایــن قانون می‌تــوان به منظور تعییــن مؤثرترین فاکتورهای کیفیت آب آشامیدنی استفاده نمود. تفسیر این قانون، به شرح زیر است:



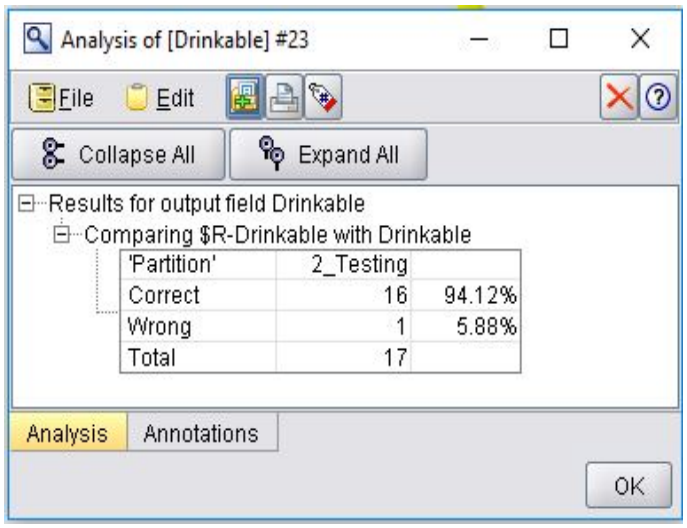
شکل ۳- تقسیم مجموعه داده‌ها به دو مجموعه داده‌های آموزشی و آزمایشی



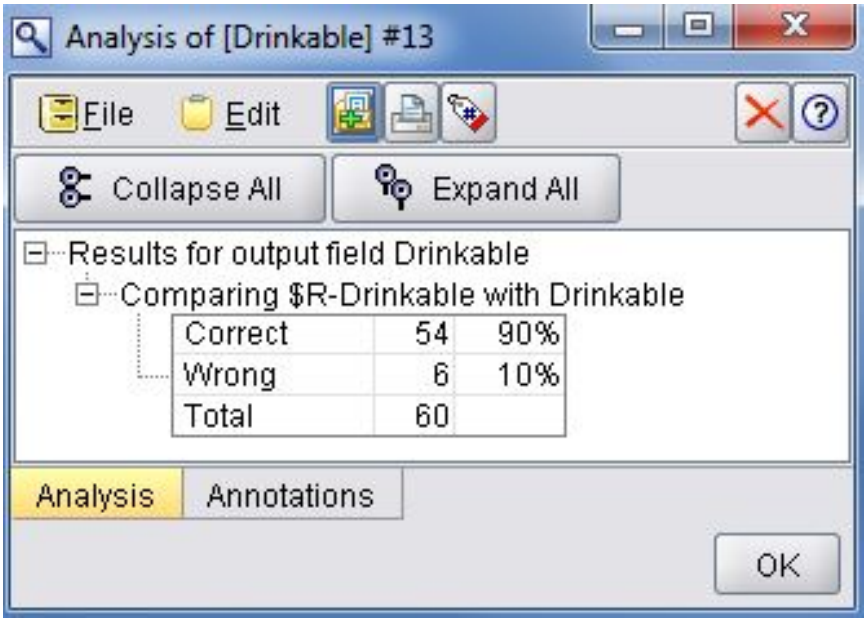
شکل ۴- قانون استخراج شده جهت تعیین مؤثرترین فاکتورهای کیفیت آب آشامیدنی

مؤثرترین فاکتورهای تاثیرگذار در کیفیت آب آشامیدنی در دشت کازرون (شکل ۴) عبارتند از: کل جامدات محلول (TDS) و کلسیم (Ca). بدین ترتیب، در صورتی‌که 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.

پس از ساخت مدل و استخراج قواعد فوق (شکل ۴)، با استفاده از الگوریتم درخت تصمیم CART، نتایج به‌دست 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)، به‌عنوان ورودی به مدل ساخته



شکل ۵- نتیجه اولیه ارزیابی انجام شده از الگوریتم CART



شکل ۶- نتیجه نهایی ارزیابی انجام شده از الگوریتم CART

شـده وارد شـد تا مشاهده شـود که مدل سـاخته شده تا چه حد قادر به تشـخیص برچسب نمونه‌های آزمایشی خواهد بود. همان‌طور که در شـکل ۵ نشان داده شده است، نتیجه ارزیابی مدل سـاخته شـده دارای اطمینـان ۹۴/۱۲ درصد اسـت. 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.

به‌منظـور اطمینان بیشـتر، کل مجموعه داده‌ها (آموزشـی و آزمایشـی) به صورت یک جا به مدل سـاخته شده وارد شد تا ارزیابـی دقیق‌تری صورت پذیرد. در این مرحله، مدل سـاخته شـده براسـاس الگوریتم درخـت تصمیـم CART موفق به تشـخیص صحیـح ۵۴ نمونـه از ۶۰ نمونه وارد شـده گردید. بنابراین همان‌طور که در شـکل ۶ نشان داده شده است، دقت به‌دسـت آمده براسـاس مؤثرترین فاکتورهای عدم کیفیت آب آشامیدنی در مدل نهایی، برابر با ۹۰ درصد است.

بحث

بـا تحلیـل مجموعـه داده‌هـا و پارامترهای انتخاب شـده و با بهره‌گیـری از تکنیـک داده‌کاوی درخـت تصمیـم و الگوریتم CART مؤثرترین فاکتورهای کیفیت آب آشامیدنی در دشت کازرون شناسـایی گردیدند. با توجه بـه 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 بیشـتر این دو عامل را می‌توان به ساختار سازندهای زمین‌شناسی منطقه و وجود کربنات کلسیم (آهک)، در ترکیب آنها مرتبط دانسـت. این نتایج مانند نتایـج پژوهش‌های Fu-Cheng و همـکار (۱۹)، Azhar و همـکاران (۲۰)، Gu و همـکاران (۲۱)، Lee و همـکار (۲۲)، Thompson (۲۳)، کاربرد تکنیک‌هـای داده‌کاوی در تعیین مؤثرترین فاکتورهای کیفیـت آب را تاییـد می‌کند. همچنین نتایـج این پژوهش با یافته‌هـای Arora و همـکاران (۱۶) ، Ji و همکاران (۱۸) ، و Soleimanpour و همـکاران (۵، ۱۰)، مبنی بر کارایی روش

CART کامـلا همخوانـی دارد؛ همچنین ایـن نتایج همانند مطالعـات 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 تقریبـا در اغلب مطالعـات حوزه کیفی منابع آب با اسـتفاده از روش‌های آماری به بررسـی رابطه بین پارامترها و بررسـی همبسـتگی بین آنها پرداخته می‌شـود؛ امـا در عالم واقعیت ممکن اسـت چندین پدیده به‌صورت همزمان بر پارامتری اثر کننـد و الگویی را به وجود بیاورند؛ بنابراین بررسـی 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 دسـت می‌دهند و به‌علت ماهیت اغلب خطی خود، از کشـف روابط غیر خطی و پیچیده بین متغیرها عاجز هستند؛ اما روش‌های داده‌کاوی به نحوی طراحی شـده‌اند که می‌توانند روابـط مرکب و پیچیده بیـن چندین پارامتـر را در پایگاه داده کشـف کنند. بنابراین روش‌های آماری، توانایی کشـف الگوهای پیچیده و غیر خطی را ندارنـد؛ در حالی‌که روش‌هـای داده‌کاوی به علت خاصیت اکتشـافی، بدون هیچ فرض اولیه‌ای شروع به مدلسازی رفتار داده‌هـا می‌نمایند و به مـرور زمان و با جلو رفتـن الگوریتم، الگـو پر رنگ‌تر و پر رنگ‌تر خواهد گردید. سـاختار غیر خطی و مقـاوم این مدل‌ها، توانایی شبیه‌سـازی رفتـار محیط‌های واقعـی (نمونه عینی آن عرصه‌هـا و محیط‌های طبیعی) را به روش‌های داده‌کاوی می‌دهد. همچنین مزیت برتر اسـتفاده از روش‌های داده‌کاوی نسـبت به روش‌هـای آماری در مطالعات منابع آب آن اسـت که روش‌های داده‌کاوی معمولا نسبت به تنظیم پارامترها حساسیت کمتری نسبت به روش‌های آماری دارنـد و دارای سـاختار انعطاف‌پذیرتری هسـتند. به‌علاوه در تکنیک‌های داده‌کاوی می‌تـوان الگویتم‌ها را به نحوی تنظیم کـرد که "خطا از حـدی کمتر" را جزء خطا به حسـاب نیاید و سـامانه تمام تلاش خود را صـرف کاهش خطاهای بزرگ و خارج از باند معرفی شده نماید؛ اما در روش‌های آماری چنین امکانی وجود نداشـته و ترم‌های خطا به سـختی قابل تغییر و سفارشی‌سازی هسـتند. بنابراین مدل‌ها بیش از اندازه کلی و غیر حساس به تغییرات هستند.

با عنایت به مباحث فوق، می‌توان بیان داشـت امروزه اسـتفاده مؤثر و اسـتخراج داده‌های پنهان از انبوه داده‌ها توسط مدیران به‌عنـوان یک هـدف عمده جهت بهبود وضعیـت موجود تلقی می‌شـود؛ و استفاده از تکنیک‌های داده‌کاوی می‌تواند به عنوان ابزار عملیاتی در شناسـایی مناطـق آلوده و ارزیابی و نظارت بر کیفیت منابع آب‌های سـطحی و زیرزمینی، و مطالعه و توسعه منابـع آب، به منظـور تصمیم‌گیری‌های علمـی برنامه‌ریزان و مدیران مفید واقع شود.

نتیجه‌گیری

با عنایت به مباحث فوق، می‌توان بیان داشـت امروزه اسـتفاده مؤثر و اسـتخراج داده‌های پنهان از انبوه داده‌ها توسط مدیران به‌عنـوان یک هـدف عمده جهت بهبود وضعیـت موجود تلقی می‌شـود؛ و استفاده از تکنیک‌های داده‌کاوی می‌تواند به عنوان ابزار عملیاتی در شناسـایی مناطـق آلوده و ارزیابی و نظارت بر کیفیت منابع آب‌های سـطحی و زیرزمینی، و مطالعه و توسعه منابـع آب، به منظـور تصمیم‌گیری‌های علمـی برنامه‌ریزان و مدیران مفید واقع شود.

توصیه می‌گـردد با توجه به کاربرد مفیـد دانش داده‌کاوی در مباحث منابع آب، نسـبت به آماربرداری‌های دقیق از این منابع و ایجـاد یک بانک اطلاعاتی جامع از آمار مسـتند از منابع آب کشـور اقدام عملی انجام شـود تـا بتوان با در اختیار داشـتن داده‌هایی صحیح و مطمئن، با اطمینان بیشـتری از این دانش جدید بهره جسـت. در پایـان با عنایت به نتایـج این پژوهش پیشـنهاد می‌گردد نسـبت به تصفیه و کاهـش املاح محلول موجود در آب توجه جدی شـود؛ همچنین بر انجام پایش‌های مستمر در قالب نمونه‌برداری‌های دوره‌ای منظم از چاه‌های این دشت تاکید می‌گردد.

ملاحظات اخلاقی

نویســـندگان کلیه نکات اخلاقی شامل عدم سرقت ادبی، انتشار دوگانـــه، تحریف داده‌ها و داده‌ســـازی را در ایـــن 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 Daneshghahi 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. Sanieeabadeh M, Mahmoodi S, Taherpour M. Applied Data Mining. Tehran: Niaze 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.

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

Original Article

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 Mesbah[1], B Hedayati[2]

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, SO_4, TH, TDS, pH, NO_3, $CaCO_3$, HCO_3, 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 $CaCO_3$ 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.