پیش بینی میزان آلودگی فلزات سنگین در رسوبات رودخانه گرگانرود با استفاده از داده کاوی

رضا طاولی[۱]، عاطفه آقاخانی*[۱]، حسین باقری

Aghakhani.atefeh@gmail.com

۱- استادیار موسسه آموزش عالی پویندگان دانش چالوس
۲- کارشناسی ارشد موسسه آموزش عالی پویندگان دانش چالوس
۳- عضو هیات علمی پژوهشگاه ملی اقیانوس شناسی و علوم جوی

چکیده

به منظور پیش بینی میزان آلودگی فلزات سنگین در رسوبات رودخانه گرگانرود با استفاده از داده کاوی، در طول رودخانه گرگان رود نمونه های رسوبی در دو فصل (بهار و تابستان) و در ۱۰ ایستگاه با سه تکرار نمونه برداری گردید. پس از آنالیز دستگاهی نمونه ها، داده های خام فلزات سنگین جمع آوری شد. سپس روش پیشنهادی مطرح گردید که شامل مراحل شروع و گردآوری داده ها، پیش پردازش داده ها ، ساخت مدل و همچنین ارزیابی و خروجی می باشد. در مرحله ساخت مدل، ساخت طبقه‌بندی با استفاده از۳ الگوریتم Naive bayes و درخت تصمیم و k-nn انجام شد و سپس ارزیابی صورت گرفت که معیار های صحت (Accuracy)، دقت (Precision)، فراخوانی(Recall) و خطا (Error) بررسی و مقایسه گردید. در خروجی روش پیشنهادی، هر ۳ الگوریتم برروی داده های مورد نظر نتایج مثبتی داشتتند. مقادیرمعیارهای صحت، دقت، فراخوانی و خطا در الگوریتم Naive bayes ، به ترتیب ٪۹۲، ٪۹۴/۴۴، ٪۸۸/۸۹، ٪۸ بدست آمد که این مقادیر درالگوریتم Naive bayes از ۲ الگوریتم درخت تصمیم و k-nn بیشتر بود. همچنین الگوریتم K-nn نسبت به درخت تصمیم خروجی بهتری داشت و مقادیرصحت و دقت در این الگوریتم بیشتر از الگوریتم درخت تصمیم بدست آمد.

واژگان کلیدی: داده کاوی ، فلزات سنگین، گرگانرود

تاریخ دریافت مقاله : ۹۸/۰۶/۱۲
تاریخ پذیرش مقاله : ۹۸/۱۲/۱۷

مقدمه

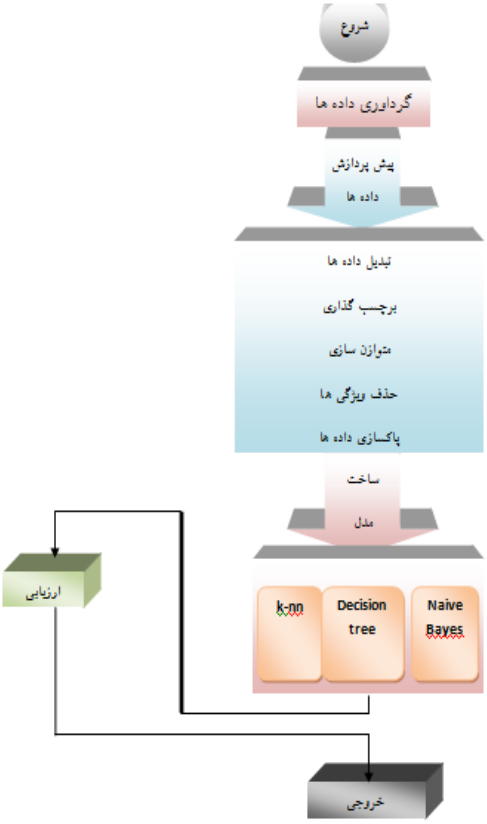
برای اولین بار مفهوم داده کاوی در کارگاه IJCAI در زمینه KDDتوسط Shapir مطرح گردید. به دنبال آن در سالهای ۱۹۹۱ تا ۱۹۹۴، کارگاه های KDD مفاهیم جدیدی را در این شاخه از علم ارائه کردند بطوری که بسیاری از علوم و مفاهیم با آن مرتبط گردیدند[۱۱]. داده کاوی یک هنر و علم استخراج اطلاعات پنهان، از مجموعه داده های فراوان است[۹]. روش ها در فرایند کشف دانش است که برای تشخیص الگوها و رابطه های نامعلوم در داده ها مورد استفاده قرار می گیرد[۱۰].

به عبارت دیگر، داده کاوی مجموعه ای از فعالیت هایی است که برای یافتن الگوهای جدید، پنهان، و غیر منتظره در داده ها استفاده می شود[۸].

داده های فراوان. این ابزارها، مدلهای آماری، الگوریتم‌های ریاضی و متدهای یادگیری ماشین (الگوریتم‌هایی که عملکرد خود را از طریق تجربه به‌صورت اتوماتیک بهبود می‌دهند) می‌باشد و فراتر از جمع‌آوری و مدیریت داده است، و شامل تجزیه و تحلیل و پیش‌گویی می باشد [۳]. داده‌کاوی یک تکنولوژی جدید نیست ولی کاربرد آن به‌طور معناداری در بخش‌های مختلف خصوصی و عمومی روبه‌رشد بوده و عموما صنایعی چون بانک، بیمه، پزشکی، خرده‌فروشی و محیط زیست از داده‌کاوی به هدف کاهش هزینه‌ها، افزایش تحقیقات و افزایش فروش استفاده می‌کنند[۴]. مفهوم کشف دانش از داده ها بیش از یک دهه است که در محیط های مالی -تجاری در حال استفاده می باشد. و در علوم مدیریت ارتباطات، مهندسی، وب کاوی، تحلیل جرایم و پزشکی جای خود را باز کرده است و به تدریج در حوزه ی زیست محیطی مورد استفاده قرار گرفت که این امر ناشی از تغییر سریع هوشیاری نسبت به اطلاعات درحوزه ی زیست محیطی است . مطالعات زیست محیطی به طور مستمر در حال تولید میزان زیادی اطلاعات و افرادی که با این نوع داد ه ها کار میکنند مواجه است، البته بین جمع آوری تا تفسیر آن داده ها شکاف وسیعی وجود دارد؛ حوزه ی به نسبت جوان و در حال رشد داده کاوی در زیست محیطی از جمله شیو ه هایی است که می تواند این صنعت را از تحلیل عمیق این داده ها بهره مند سازد و به توسعه تحقیقات زیست محیطی و تصمیم گیری های علمی دراین زمینه منتج شود . داده کاوی به کندی اما به طور فزاینده ای برای رفع مشکلات متعدد در کشف دانش دربخش زیست محیطی به کارگرفته شده است [۱،۲،۵]. از مهم ترین دلایل رشد کند این علم در این بخش حساسیت علم زیست محیطی و گره خوردن آن با سلامتی انسان ها است و مهم ترین چالش این است که اگر فرض بر این باشد که نتایج داده کاوی به طور کامل قابل اعتماد است؛ تغییر عادت ارایه دهندگان تحقیقات زیست محیطی سنتی به زیست محیطی مبتنی بر شواهد دشوار است. با این وجود، امروزه بخش زیست محیطی نیازیشتری به داده کاوی پیدا کرده است و حرکت از مطالعات زیست محیطی سنتی به سمت زیست محیطی مبتنی بر شواهد از جمله مواردی است که میتواند مؤکد این امر باشد [۶]. در این پژوهش سعی می گردد تا ضمن بررسی میزان آلودگی فلزات سنگین در رسوبات رودخانه گرگانرود با استفاده از داده کاوی، شیوه ای جدید را جهت پیش بینی میزان آلودگی ها ارایه گردد.

روش کار

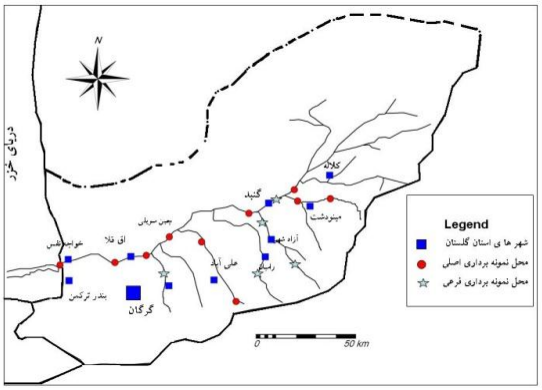
روش پیشنهادی انجام پژوهش شامل مراحل شروع و گردآوری داده ها، پیش پردازش داده ، ساخت مدل و همچنین ارزیابی و خروجی می باشد (شکل ۱).



شکل(۱) فلوچارت روش پیشنهادی

گرد آوری اطلاعات

در این پژوهش به منظور گردآوری اطلاعات جهت پیش بینی میزان آلودگی فلزات سنگین در رسوبات رودخانه گرگانرود با استفاده از داده کاوی، ابتدا در طول رودخانه نمونه های رسوبی در دو فصل (بهار و تابستان) در ۱۰ ایستگاه با سه تکرار برداشت گردید (شکل۱). پس از ارسال نمونه ها به آزمایشگاه و آنالیز دستگاهی نمونه ها، داده های خام فلزات سنگین گرد آوری شد. جامعه آماری شامل ۶۰ نمونه حاصل ۱۰ ایستگاه با ۳ تکرار و دو فصل بهار و تابستان است و ۱۱ ویژگی را شامل می شود.



شکل(۲) نقشه موقعیت برداشت نمونه ها

پیش پردازش داده

مرحله دوم پیش پردازش داده ها می باشد که ازاهمیت زیادی برخورداراست. در این مرحله، اماده سازی داده ها صورت می گیرد. به ترتیب عملیات تبدیل داده ها، برچسب گذاری، متوازن سازی داده ها و حذف ویژگی ها انجام گرفت (شکل۳).

Row No.	pollution	Element	اثر آلودگی As	اثر آلودگی Fe	اثر آلودگی Ni	اثر آلودگی Cd	اثر آلودگی Mn	اثر آلودگی Zn	Hg
1	yes	S1	severe	severe	medium	low	medium	low	low
2	yes	S1	medium	severe	medium	low	medium	low	low
3	yes	S1	medium	severe	medium	low	medium	low	low
4	no	S2	medium	severe	medium	low	medium	low	low
5	no	S2	medium	severe	medium	low	medium	low	low
6	no	S2	medium	severe	medium	low	medium	low	low
7	no	S3	medium	severe	medium	low	medium	low	low
8	no	S3	medium	severe	medium	low	medium	low	low
9	no	S3	medium	severe	medium	low	medium	low	low
10	no	S4	medium	severe	medium	low	medium	low	low
11	no	S4	medium	severe	medium	low	medium	low	low
12	no	S4	medium	severe	medium	low	medium	low	low
13	no	S5	medium	severe	medium	low	medium	low	low
14	no	S5	medium	severe	medium	low	medium	low	low
15	no	S5	medium	severe	medium	low	low	low	low

شکل(۳) نمونه ای از داده های پیش پردازش شده

تبدیل داده (data transformation)

هر الگوریتم داده کاوی بر اساس نوع خروجی و هدفی که دنبال می‌کند به فرمت خاص خود نیاز دارد. در این مرحله باید داده‌های مورد نیاز الگوریتم را به شکل و قالب قابل قبول برای الگوریتم تبدیل کنیم که ما در این پایان نامه داده های عددی را به داده های اسمی تبدیل میکنیم. برای سهولت کارو برای اینکه از کار نتیجه درستی داشته باشیم داده های عددی را به داده های اسمی تبدیل کردیم. میدانیم برای هر فلز سنگین یک مقداراستاندارد جهانی تعریف شده است [۷].که غلظت مقدار فلزات سنگین جهان از جمله رودخانه گرگانرود با مقدار استانداردش مقایسه می شود. بدین صورت که مقدار داده ی عددی را بر مقدار استاندارد جهانی ان تقسیم نمودیم.بازه عددی (۰،۱) رادرنظر بگیرید ،اگر مقدار بدستآمده صفرباشد ان ویژگی را ضعیف (low) واگر ۱باشد شدید (severe) نامگذاری می گردد.اگر مابین صفر تا یک بدستآمد متوسط (medium) نامگذاری می شود.بنابراین ۱۰۵ نمونه داریم که همان ایستگاه ها میباشند و ۱۱ویژگی با نام اثرآلودگی فلزات سنگین داریم که هرویژگی شامل ۳ مقدار: ضعیف ، متوسط و شدید است

برچسب گذاری داده

اجرای این مرحله ازتحقیق هم مهم میباشد که نباید درانجام ان کوتاهی نمود. بر روی ویژگی هدف برچسب گذاری (label) انجام دادیم. ویژگی هدف در این پایان نامه، آلودگی(Pollution) میباشد.بنابراین بابرچسب گذاری برروی ویژگی آلودگی، این ویژگی را به عنوان ویژگی هدف مشخص گردید.

متوازن سازی

داده های این تحقیق نامتوازن بودند.یکی از روش های مواجهه با مجموعه داده های نامتوازن، بکارگیری روش over sampling است که بسیار کارامد است. بنابراین با بکارگیری روش over sampling ،داده ها به داده های متوازن تبدیل شد و با ۵ بارتکرارکردن صفت yes داده هارا متوازن گردید. از تعداد ۶۲ داده به تعداد۱۰۶ داده رسید.

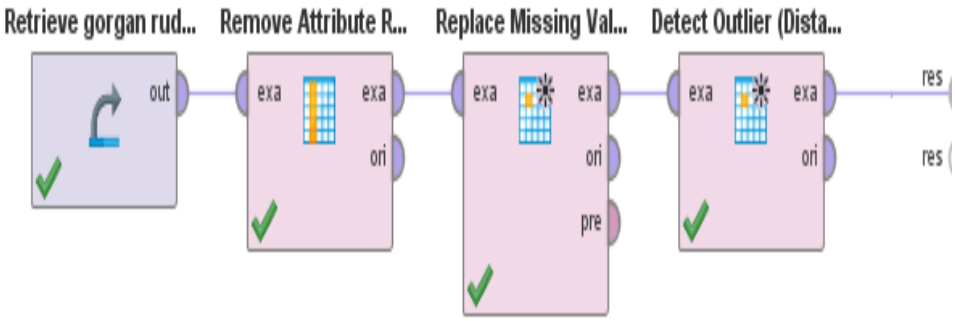
حذف ویژگی

یکی از مهمترین نکات در داده کاوی این است که ممکن است همیشه، همه داده ها مورد نیاز نباشند و تنها بخشی از داده ها که مورد نیاز است باید مورد پردازش قرار بگیرد. عملگر remove attribute به این مباحث می پردازد. با استفاده ازعملگر remove attribute ستون ۸ که همان ویژگی اثار

آلودگی جیوه است را حذف گردید زیرا این داده ها در نتیجه کار تأثیری ندارد و مقدار ان زیر حد تشخیص دستگاه بوده است.

پاکسازی داده(data cleaning)

هر چه این گام از داده کاوی بهتر انجام شود، خروجی الگوریتم ها و تکنیک های داده کاوی کیفیت بالاتری خواهد داشت.برای انجام این کار، از عملگر missing value , detect outlier استفاده کردیم.درنهایت همانطورکه در شکل(۴) مشاهده می کنید عمیات پیش پردازش داده ها انجام گرفت.



شکل(۴) عملیات پیش پردازش داده

ساخت مدل

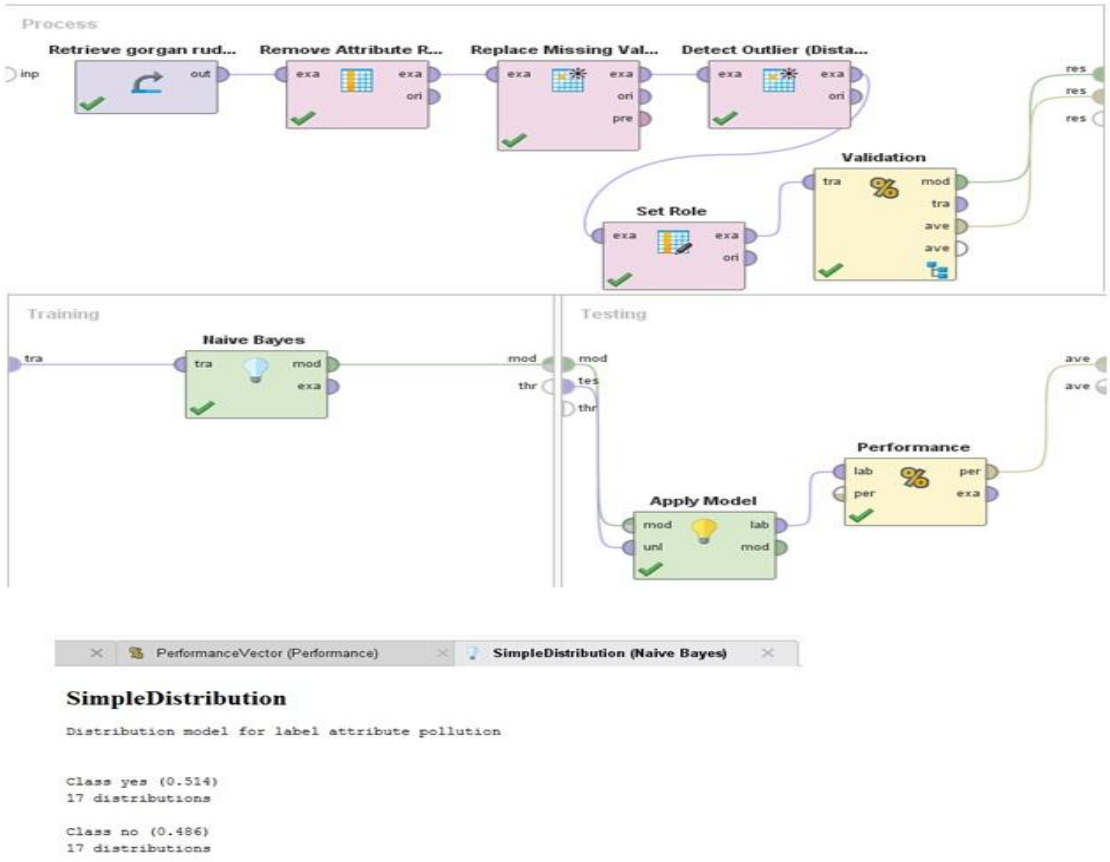
مرحله ساخت مدل با استفاده از الگوریتم‌های متنوع و با در نظر گرفتن ماهیت داده، نظم‌های مختلف موجود در داده‌ها شناسائی می‌شود. بطور کلی روش‌های مختلف کاوش داده را به دو گروه روش‌های پیش بینی و روش‌های توصیفی طبقه بندی می‌کنند که دراین پژوهش از روش‌های دسته بندی استفاده شد.

دسته بندی (Classification)

در روش‌های دسته بندی(الگوریتم‌های دسته بندی) مجموعه داده اولیه به دو مجموعه داده با عنوان مجموعه داده‌های آموزشی (Train Dataset) و مجموعه داده‌های آزمایشی (Test Dataset)تقسیم می‌شود. می‌دانیم هر Case شامل مجموعه ای از Attribute (ویژگی) هاست، که یکی از این ویژگی‌ها، ویژگی دسته (ویژگی هدف) نامیده می‌شود. در مرحله آموزش، مجموعه داده‌های آموزشی به یکی از الگوریتم‌های دسته بندی داده می‌شود تا بر اساس سایر ویژگی‌ها برای مقادیر ویژگی دسته، مدل ساخته شود.پس از ساخت مدل، در مرحله ارزیابی؛ دقت مدل ساخته شده به کمک مجموعه داده‌های آزمایشی ارزیابی خواهد شد. در الگوریتم‌های دسته بندی از آنجا که ویژگی دسته مربوط به هر Case مشخص است به صورت الگوریتم‌های با ناظر محسوب می‌شوند. بدیهی است که تشخیص بر اساس دسته هایی است که مدل در مرحله آموزش با آنها روبرو شده است؛ بنابراین امکان تشخیص دسته جدید در کاربرد دسته بندی وجود نخواهد داشت [۹]. بااستفاده از۳ الگوریتم Naive bayes، درخت تصمیم و k-nn دسته بندی داده ها انجام گرفت و در نهایت ارزیابی معیار های صحت دقت (Accuracy) و دقت(precision) بررسی و مقایسه گردید.

نتایج

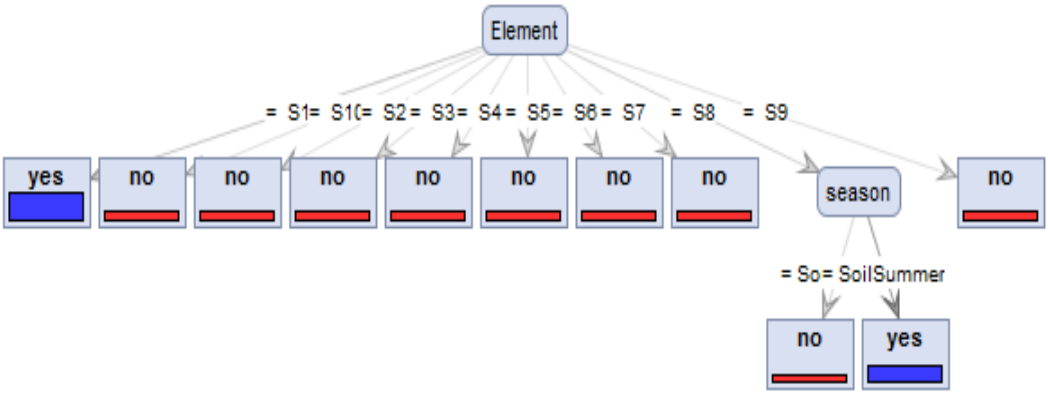
به طور ساده روش Naive Bayes روشی برای دسته بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده‌است. ابتدا ساخت مدل بدون ارزیابی در الگوریتم Naive bayes صورت گرفت و مقادیر صحت ،دقت، فراخوانی و خطا به ترتیب مقادیر ٪۹۷/۱۴،٪۹۷/۳۷،٪۹۷/۰۶،٪۲/۸۶ بدستآمد در حالی خروجی الگوریتم Naive bayes با ارزیابی را برای معیارهای فوق ترتیب مقادیر ٪۹۲ ،٪۹۴/۴۴ ، ٪۸۸/۸۹، ٪۸ بدست آمد. باتحلیل جواب الگوریتم Naive bayes نشان می دهد که برای هردو کلاس بله و خیر،۱۷ توزیع داریم. مدل توزیع برای صفت برچسب آلودگی (شکل ۵).کلاس بله (۰،۵۱۴) و ۱۷ توزیع، کلاس بدون (۰،۴۸۶) ۱۷و توزیع.



شکل(۵) اجرای الگوریتم Naive bayes با ارزیابی

درخت تصمیم(Decision tree) یک الگوریتم با نظارت است و همانند ماشین بردار پشتیبان برای کلاس بندی استفاده می شود. درخت تصمیم از یک درخت برای ساخت یک مدل پیش بینی (تخمین) استفاده می کند که مشاهدات درباره یک آیتم را به نتیجه گیری هایی درباره مقدار هدف آن آیتم نگاشت می دهد.

خروجی درخت تصمیم بدون ارزیابی را برای معیارهای صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۱۰۰، ٪۱۰۰، ٪۱۰۰، ٪۰ بدست آمد و در قدم بعدی درخت تصمیم را با ارزیابی (validation) انجام گرفت شکل(۶) خروجی درخت تصمیم با ارزیابی را برای معیارهای صحت ، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۸۰/۶۵، ٪۸۶/۳۶، ٪۸۰، ٪۱۹/۳۵، بدست آمد.



شکل(۶) خروجی الگوریتم درخت تصمیم با ارزیابی

همانگونه که در شکل پیداست مقدار ویژگی هدف yes از no کمتر دیده می شود و ۲ فصل بهار و تابستان را به صورت جداگانه نشان می دهد. جستجوی نزدیک ترین همسایه یا Nearest Neighbor (KNN)، که همچنین با نامهای جستجوی مجاورت، جستجوی همسانی یا جستجوی نزدیک ترین نقطه شناخته می‌شود، یک مسئله بهینه سازی برای پیدا کردن نزدیک ترین نقطه‌ها در فضاهای متریک است استفاده از الگوریتم KNN نیازمند تعیین سه موضوع می‌باشد :

باید یک مجموعه رکورد داشته باشیم، یک معیار محاسبه شباهت نیز باید داشته باشیم و -مقدار K نیز باید مشخص شود تا بتوان بر اساس آن عمل نمود. انتخاب مقدار Kدر این روش دسته‌بندی بسیار مهم و کلیدی است. اگر مقدار K خیلی کوچک انتخاب شود، الگوریتم به نویز حساس می‌شود. در واقع نویزها نزدیک آن رکورد ممکن است ایجاد اشتباه کنند. اگر مقدار Kخیلی بزرگ انتخاب شود، ممکن است در میان نزدیک‌ترین همسایه‌ها، رکوردهایی از دسته‌های دیگر نیز قرار بگیرند.وقتی K عدد بزرگی انتخاب شود، منجر به خطای دسته‌بند در دسته‌بندی رکورد ورودی خواهد شد. یکی از ایده‌هایی که برای حل این مشکل ارائه شده، تعریف فاکتور وزن است. این فاکتور وزنی برابر d2/1را در نظر می‌گیرد که مقدار d بیانگر فاصله هر رکورد تا رکورد ورودی می‌باشد. به این ترتیب فاصله‌ها برای الگوریتم اهمیت پیدا می‌کنند و این وزن‌دهی سبب می‌شود که به رکوردهایی که نزدیک‌تر به رکورد ورودی هستند اهمیت بیش‌تری داده شود [۸].

براین اساس پیاده سازی الگوریتم K-NN بدون ارزیابی برای K اعداد فرد را جایگذاری میکنیم.

اگرk=1 باشد، مقدار صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۱۰۰، ٪۱۰۰، ٪۱۰۰، ٪۰ بدست می آید و اگرk=3 باشد، مقدار صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۱۰۰، ٪۱۰۰، ٪۱۰۰، ٪۰ حاصل می شود و اگر k=5 باشد، مقدار صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۹۸/۱۰ ، ٪۹۸/۲۱، ٪۹۸/۰۴، ٪۱/۹۰ بدست می آید و نهایتا اگرk=7 باشد، مقدار صحت، دقت، فراخوانی و خطا را نشان می دهد که به ترتیب مقادیر ٪۱۰۰، ٪۱۰۰، ٪۱۰۰، ٪۰ بدست آمد.

این الگوریتم با ارزیابی (validation) به صورت زیر انجام گرفت .

اگرk=1 باشد، مقدار صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۸۰/۶۵، ٪۸۶/۳۶، ٪۸۰، ٪۱۹/۳۵بدست آمد. مدل نزدیکترین همسایه برای طبقه بندی این مدل شامل ۱۰۵ نمونه با ۱۷ بعد از کلاس بله و خیر است:

اگرk=3 باشد، مقدار صحت، دقت، فراخوانی به ترتیب مقادیر ٪۸۰/۶۵، ٪۸۶/۳۶، ٪۸۰، ٪۱۹/۳۵ بدست آمد و مدل نزدیک ترین همسایه برای طبقه بندی این مدل شامل ۱۰۵ نمونه با ۱۷ بعد از کلاس بله و خیر است:

اگرk=5باشد، مقدار صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۹۰/۳۲، ٪۹۲/۱۱، ٪۹۰، ٪۹/۶۸ بدستآمد.جواب تحلیل k-nn در k=5 به صورت مدل نزدیکترین همسایه برای طبقه بندی شامل ۱۰۵ نمونه با ۱۷ بعد از کلاس بله و خیر است:

اگرk=7 باشد، مقدار صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۹۰/۳۲، ٪۹۲/۱۱، ٪۹۰، ٪۹/۶۸ بدست آمد.

.اگر k=9 باشد، مقدار صحت، دقت، فراخوانی و خطا به ترتیب مقادیر ٪۱۰۰، ٪۱۰۰، ٪۱۰۰، ٪۰ بدستآمد.بدین ترتیب برای همه ی k ها ۱۷ بعد بدست آمد. با توجه به نتایج بدست آمده، بعد از پیاده سازی الگوریتم K-NN با ارزیابی بیشترین مقدار برای accuracy و precision در K= ۵ مشاهده گردید، بنابراین برای الگوریتم k-nn مقادیر accuracy و precision در k= ۵ را انتخاب میکنیم.

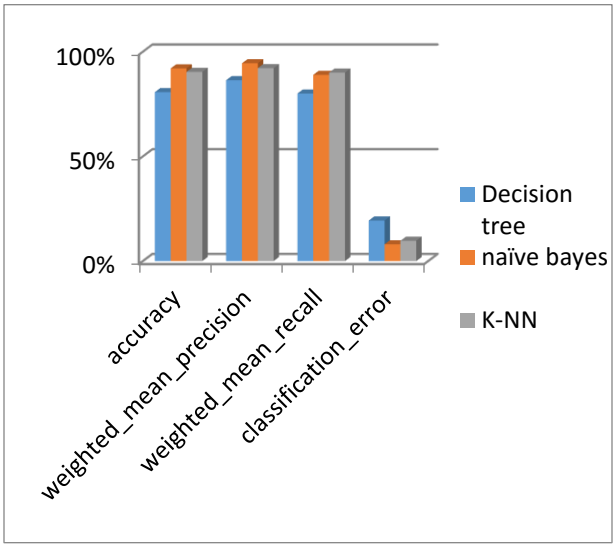
بحث و نتیجه گیری

مقادیرمعیارهای صحت، دقت، فراخوانی و خطا را برای ۳ الگوریتم Naive bayes، درخت تصمیم و k-nn به ترتیب در ساخت مدل بدون ارزیابی و با ارزیابی به صورت جدول و نمودار ترسیم گردید. در نهایت جدول مقایسه ای ۳ الگوریتم ذکر شده با ارزیابی ترسیم شد .در واقع جدول و نمودار فوق الذکر از بهترین جواب این الگوریتم ها حاصل شده است. در جدول ۱ و شکل(۷)، نتایج ۳ الگوریتم Naive bayes، درخت تصمیم و k-nn را مقایسه کردیم و نتایج کلی زیر بدست آمد.

هر ۳ الگوریتم برروی داده های ما نتایج مثبتی داشتند.مقادیر Accuracy و precision در Naive bayes ازمقادیر این دو معیار در ۲ الگوریتم درخت تصمیم و k-nn بیشتر است. مقادیرمعیارهای صحت، دقت، فراخوانی و خطادر الگوریتم نیوبیز، به ترتیب ٪۹۲، ٪۹۴/۴۴، ٪۸۸/۸۹، ٪۸ بدست آمد. بنابراین در این پژوهش با این داده ها الگوریتم Naive bayes جواب بهتری میدهد. همچنین K-nn هم نسبت به درخت تصمیم جواب بهتری داد زیرا مقادیر Accuracy وprecision در k-nn بیشتر از درخت تصمیم است. برطبق درخت تصمیم، مشخص است که تعداد ایستگاه هایی که آلودگی دارند (کلاس yes) خیلی کمتر از تعداد ایستگاه هایی است که آلودگی ندارند (کلاس No). در فصل بهاردر ایستگاه S1 (خواجه نفس) درفصل تابستان در ایستگاه S8 که همان ایستگاه گنبد(ورودی) است، آلودگی وجود داشت ولی در ایستگاه های دیگر آلودگی دیده نشد. نتایج این پژوهش همانند نتایج Cho [۱۳] و Rajakumari و همکاران [۱۲] موید اثر بخشی این روش در مطالعات منابع آب و کنترل کیفی زیست محیطی می باشد.

جدول ۱) نتیجه۳ الگوریتم naïve bayes،درخت تصمیم و K-NN با ارزیابی

with validation	Decision tree	naïve bayes	K-NN
accuracy	80.65%	92.00%	90.32%
weighted_mean_precision	86.36%	94.44%	92.11%
weighted_mean_recall	80.00%	88.89%	90%
classification error	19.35%	8.00%	9.68%



شکل(۷) نتیجه۳ الگوریتم naïve bayes،درخت تصمیم و K-NN با ارزیابی

منابع

[۱] ستاری، م.، نایب زاده، ع.، نجف آبادی میرعباسی، ر.،۱۳۹۳. پیش بینی کیفیت آب‌های سطحی با استفاده از روش درخت تصمیم. مجله مهندسی آبیاری و آب. دوره۴ شماره۱۵ ص.۷۶-۸۸

[۲] حاجیان نژاد، م.، رهسپار.، ۱۳۸۹.بررسی تاثیر روان آب ها وپساب تصفیه خانه فاضلاب بر پارامترهای کیفی آب رودخانه زاینده رود. مجله تحقیقات نظام سلامت /سال ششم/ویژه نامه، ص.۸۲۱-۸۲۸.

[۳] مهریزی، حائری، علی اصغر ، «داده‌کاوی: مفاهیم، روش‌ها و کاربردها» (۱۳۸۲) پایان‌نامه کارشناسی ارشد آمار اقتصادی و اجتماعی، دانشکده اقتصاد، دانشگاه علامه طباطبائی.

[۴] افراز، علی، مدیریت دانش (مفاهیم، مدل‌ها، اندازه‌گیری و پیاده‌سازی)، چاپ اول، انتشارات دانشگاه صنعتی امیرکبیر، ۱۳۸۴.

[۵] رضوانی گیل کلایی، سهراب، بررسی آثار هیستوپاتولوژیک ناشی از برخی آلاینده های زیست محیطی دریای خزر روی ماهیان استخوانی شکارچی آزاد و سوف دریای خزر، ۱۳۸۶.

[۶] دبیری، مجتبی. ۱۳۸۲. آلودگی محیط زیست، هوا- آب- خاک- صوت. گروه شیمی دانشکده علوم دانشگاه شهید بهشتی. تهران، نشر اتحاد.۳۹۹ص.

[۷] شریعت فیض آبادی، فاطمه، استاندارد های کیفی آب، ۱۳۷۷.

[8] 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.

[9] Shahrabi J. Data Mining. Tehran: Jahad-e Daneshghahi Publication; 2013 (in Persian)

[10] Guilford .P, Kline. R.B., 2005.Data PreParation and Screening, in Principles and Practice of Structural Equation Modeling, Chapter 3.pp.45-62

[11] Mehmed Kantardzic , John Wiley & Sons.,2003.Data Mining: Concepts, Models, Methods, and Algorithms, Preparing the Data,Chapter 2,3.

[12] Rajakumari SB, Nalini C. Identification of leadcontaminant in river water quality data. Journal of Chemical and Pharmaceutical Sciences.2016;9(4):2764-66.

[13] 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.