doi.org/10.22093/wwj.2020.209875.2958

Journal of Water and Wastewater, Vol. 31, No.4, pp: 184-193

Modeling of Unauthorized Water Consumption Detection (Case Study: Qom)

Gh. Amini

MSc of Statistics, Water and Wastewater Co., Qom Province, Qom, Iran
Ghasem2094.amini@gmail.com

(Received Nov. 30, 2019 Accepted Feb. 12, 2020)

To cite this article:
Amini, Gh. 2020. "Modeling of unauthorized water consumption detection (case study: Qom)"
Journal of Water and Wastewater, 31(4), 184-193. Doi: 10.22093/wwj.2020.209875.2958. (In Persian)

Abstract

Detection of tampering in water meters as part of unauthorized usage is a key step in development of service delivery and increasing water resource productivity, and requires special attention. Data used to identify unauthorized water usage, due to tampering in water meters, include 671 subscribers with a history of meter tampering during the years 2017-2019 and a random sample of 3120 subscribers with no tampering record (clean) among Qom's residential water users. Data analysis was conducted using subscriber's water consumption and invoice payment history as well as supervised data mining techniques such as decision tree, support vector machine, neural network, logistic regression, K-nearest neighbor and unsupervised clustering method. The comparison of different data mining techniques between two groups of tampered and non-tampered water meters showed that among the supervised methods, the accuracy of the models is close to each other and there is a 1–3% difference between them. On the other hand, given the percentage of correct responses among the methods, logistic regression, as the best data mining model, with correct detection of 85% of tampered and 91% of non-tampered cases as well as 89% overall accuracy on the testing data, can be used for identification of tampered meters. The study used clustering as an unsupervised technique. The subscribers were grouped into six clusters. Cluster 3 (n=160 subscribers) showed distinct behavior from the other clusters. About 86% of subscriptions in cluster 3 are tampered cases. Moreover, 18% of the tampered cases detected by logistic regression are in this cluster. Data mining techniques for identification of water meter tampering were presented in this study. Findings of the study indicated that both supervised (including response variable) and unsupervised methods (no response variable) such as clustering can be used for the identification of unauthorized water consumption In this study, logistic regression, due to its high accuracy, was selected as the most appropriate model for detection of tampered meters.

Keywords: Data Mining, Unauthorized Usage, Meter Tampering, Logistic Regression, Clustering.

doi.org/10.22093/wwj.2020.209875.2958

Original Paper

مجله آب و فاضلاب، دوره ۳۱، شماره ۴، صفحه: ۱۹۳-۱۸۴

مدل‌سازی تشخیص مصرف غیرمجاز آب (مطالعه موردی: شهر قم)

قاسم امینی

۱- کارشناس ارشد آمار شرکت آب و فاضلاب شهری استان قم، قم، ایران
Ghasem2094.amini@gmail.com

(دریافت ۹۸/۹/۹ پذیرش ۹۸/۱۱/۲۳)

برای ارجاع به این مقاله به صورت زیر اقدام بفرمایید:
امینی، ق.، ۱۳۹۹، " مدل‌سازی تشخیص مصرف غیرمجاز آب (مطالعه موردی: شهر قم)"مجله آب و فاضلاب، ۳۱(۴)، ۱۹۳-۱۸۴.
Doi: 10.22093/wwj.2020.209875.2958

چکیده

شناسایی کنتورهای دست‌کاری شده آب، به‌عنوان بخشی از مصارف غیرمجاز یکی از گام‌های اساسی در توسعه خدمات‌رسانی و افزایش بهره‌وری از منابع آب به حساب می‌آید که باید به آن توجه کرد. داده‌های مورد استفاده برای شناسایی مصرف غیرمجاز حاصل از دست‌کاری کنتور شامل تعداد ۶۷۱ مشترک با سابقه دست‌کاری کنتور در سال‌های ۹۸-۹۶ و یک نمونه تصادفی ۳۱۲۰ تایی از مشترکان بدون سابقه دست‌کاری (سالم) در کاربری خانگی شهر قم است. به‌منظور تحلیل داده‌ها از سابقه مصرف و پرداخت صورت‌حساب مشترکان و همچنین روش‌های با نظارت داده کاوی از قبیل درخت تصمیم، ماشین‌بردار پشتیبان، شبکه عصبی، رگرسیون لجستیک، نزدیک‌ترین همسایگی و روش بدون نظارت خوشه‌بندی استفاده شد. مقایسه روش‌های مختلف داده کاوی بین دو گروه کنتورهای دست‌کاری شده و سالم نشان داد در بین روش‌های با نظارت، دقت مدل‌ها نزدیک یکدیگر است و اختلاف ۱ تا ۳ درصدی بین آنها وجود دارد. از طرفی با توجه به درصد پاسخ صحیح در بین روش‌ها، رگرسیون لجستیک با تشخیص صحیح ۸۵ درصد موارد دست‌کاری شده و ۹۱ درصد موارد سالم و دقت کلی ۸۹ درصد بر روی داده‌های تست به‌عنوان بهترین مدل برای شناسایی کنتورهای دست‌کاری شده می‌تواند استفاده شود. همچنین با استفاده از خوشه‌بندی به‌عنوان یک روش بدون نظارت، مشترکان در ۶ خوشه دسته‌بندی و خوشه شماره ۳ با تعداد ۱۶۰ مشترک، رفتار مجزایی از سایر خوشه‌ها نشان داد. ۸۶ درصد اشتراک‌های این خوشه شامل موارد دست‌کاری شده بودند و ۱۸ درصد از موارد دست‌کاری تشخیص داده شده توسط رگرسیون لجستیک در این خوشه قرار گرفت. این پژوهش به ارائه مدل‌هایی برای شناسایی موارد دست‌کاری کنتور آب توسط تکنیک‌های داده کاوی پرداخته است. با توجه به یافته‌های پژوهش، به‌منظور شناسایی مصارف غیرمجاز آب، می‌توان از هر دو روش با نظارت (شامل متغیر پاسخ) و بدون نظارت (بدون نیاز به متغیر پاسخ) مانند خوشه‌بندی استفاده کرد. در این پژوهش رگرسیون لجستیک با دقت زیاد به‌عنوان مناسب‌ترین مدل برای شناسایی کنتورهای دست‌کاری شده انتخاب شد.

واژه‌های کلیدی: داده کاوی، مصرف غیرمجاز، دست‌کاری کنتور، رگرسیون لجستیک، خوشه‌بندی

۱- مقدمه

کشور ایران، به‌دلیل استقرار در کمربند خشک جغرافیایی و نوار بیابانی، در زمره مناطق کم باران جهان قرار گرفته است، به‌طوری که بیشتر اراضی در ایران مرکزی مانند استان قم از یک بارندگی مؤثر محروم است. از طرفی این استان با توجه به موقعیت خاص استراتژیکی، سیاسی و مذهبی یکی از استان‌هایی است که تمرکز جمعیت در آن به‌شدت رو به افزایش است و همین موضوع باعث

doi.org/10.22093/wwj.2020.209875.2958

رشد صنعت و بالا رفتن سطح زندگی مردم و در نتیجه تقاضای آب و نیاز شدید به آن شده است (Amini and Saeidi, 2017).

از طرفی اهمیت حفظ کیفیت آب شرب و نیز محدود بودن منابع آبی در دسترس همواره یکی از دغدغه‌های اصلی برنامه‌ریزان صنعت آب در کشور است. با توجه به شرایط کنونی بحران آب و افزایش تقاضا برای این ماده حیاتی و کالای ارزشمند، نگاه فراگیر به آب و مدیریت آن اهمیت خاصی دارد (Amini et al., 2018a). شرکت‌های آب و فاضلاب با نگاه ویژه به این موضوع، همواره برای جلب رضایت مشترکان، یعنی تأمین و توزیع آب با قیمت مناسب، حداقل کردن هزینه‌ها و مدیریت مصرف گام برداشته‌اند (Amini et al., 2018b). این مهم هنگامی میسر خواهد شد که این شرکت‌ها شناخت و درک درستی نسبت به رفتار مصرفی مشترکان خود داشته باشند. امروزه با توجه به گسترش فناوری اطلاعات و توسعه سیستم‌های اطلاعاتی در سازمان‌ها، با استفاده از ابزارها و الگوریتم‌های داده کاوی[1] می‌توان ماهیت پیچیده رفتار مصرفی مشترکان را در قالب الگوهایی شناسایی و مصرف را مدیریت بهینه کرد.

یکی از مصادیق مهم مدیریت مصرف و افزایش بهره‌وری از منابع آب، شناسایی مصارف غیرمجاز آب است. موضوعی که مسائل و مشکلات متعددی را برای صنعت آب و فاضلاب به وجود آورده است و خسارات مستقیم و غیرمستقیم ناشی از آن باعث شده طی سال‌های اخیر شرکت‌های آب و فاضلاب نگاه ویژه‌ای به این‌گونه انشعابات داشته باشند و برخورد با عاملان را با انسجام بیشتری، پیگیری کردند و با اجرای طرح‌های مقابله با انشعابات غیرمجاز آب، نقش مؤثری در کاهش این پدیده ناهنجار داشته باشند (Amini et al., 2018a).

مشترکان با مصارف غیرمجاز را می‌توان به دو دسته تقسیم کرد، دسته اول که مبنای این پژوهش است مشترکانی هستند که کنتور آب دارند ولی بنا بر دلایلی اقدام به دستکاری کنتور می‌کنند. دسته دوم، فاقد اشتراک است و کنتور آب ندارند و مستقیم از شبکه توزیع استفاده می‌کنند که باعث اخلال در مدیریت توزیع، تخریب شبکه آب‌رسانی، احتمال ایجاد آلودگی و نشت آن به شبکه در هنگام نصب انشعاب و هدر رفت آب به‌علت استفاده از اتصالات و لوله‌های غیر استاندارد، افت فشار و در نتیجه نارضایتی مردم خواهد شد (Amini et al., 2018a).

برخی پژوهش‌های انجام شده توسط داده‌کاوی در خصوص کشف تقلب در ادامه آورده شده است. مینایی و همکاران در سال ۲۰۱۱ تشخیص کلاهبرداری در سازمان‌های خدمات‌رسانی را با داده‌کاوی انجام دادند (Minaie et al., 2011). هاشم و حمید در سال ۲۰۱۲ با استفاده از داده‌کاوی و روش‌های نزدیک‌ترین همسایگی، شبکه عصبی و ماشین‌بردار پشتیبان در سه مجموعه داده سالانه، فصلی و ماهانه به شناسایی موارد دستکاری شده در کنتور آب پرداختند. در هر سه مجموعه داده، ماشین‌بردار پشتیبان دقت بیشتری نسبت به روش‌های دیگر دارد و بهترین نتیجه با دقت ۹۳ درصد مربوط به داده‌های فصلی حاصل شد (Hashem and Humaid, 2012).

بررسی و مقایسه روش‌های مختلف داده‌کاوی در تشخیص تقلب توسط آنیتا و راویندرا انجام شد (Anita and Ravindra, 2013). موندرو و همکاران در سال ۲۰۱۵ با استفاده از سه الگوریتم به شناسایی کنتورهای دستکاری شده آب در یکی از روستاهای اسپانیا پرداختند. آنها با استفاده از این الگوریتم‌ها، ۸۵ اشتراک را به‌عنوان نمونه‌هایی با احتمالی دستکاری زیاد انتخاب و پس از پیمایش توسط بازرسان ملاحظه کردند تنها ۷ درصد آنها دستکاری کنتور داشته‌اند (Monedero et al., 2015).

بررسی روش‌های تشخیص مصرف غیرمجاز برق بر اساس داده‌های اندازه‌گیری در ساختار شبکه هوشمند (Kasaeyan and Ghayni, 2017) و استفاده از تکنیک‌های ماشین‌بردار پشتیبان، درخت تصمیم، رگرسیون لجستیک در تشخیص تقلب کارت‌های اعتباری از دیگر پژوهش‌های انجام شده است (Navanshu and Saad, 2018, Monika and Amarpreet, 2018).

امینی و همکاران در سال ۲۰۱۸ با استفاده از خوشه‌بندی دو مرحله‌ای و استفاده از ۱۴ دوره میانگین مصرف ماهانه ۱۱۷۰۱۷ مشترک خانگی شهر قم در سال‌های ۹۵ تا ۹۶، مشترکان را به سه خوشه دسته‌بندی و یکی از خوشه‌ها با تعداد ۳۵۸۷ مشترک را به‌عنوان نمونه احتمالی برای مصارف دستکاری کنتور شناسایی کردند (Amini et al., 2018a).

متأسفانه در داخل کشور پژوهش جامعی در خصوص شناسایی مصارف غیرمجاز آب حاصل از دستکاری کنتور توسط روش‌های

[1] Data mining

doi.org/10.22093/wwj.2020.209875.2958

مدل‌سازی انجام نشده است. شرکت‌های آب و فاضلاب نیز معمولاً شناسایی موارد دستکاری کنتور را به‌صورت سنتی مانند گزارش مأمورین قرائت انجام می‌دهند. از طرفی در پژوهش‌های محدودی که در رابطه با دستکاری کنتور وجود دارد، بیشتر از روش‌های داده‌کاوی با نظارت استفاده شده است و از روش بدون نظارت مانند خوشه‌بندی استفاده نشده است. بنابراین هدف این پژوهش ارائه روشی علمی، جامع و مفید با استفاده هم‌زمان از الگوریتم‌های با نظارت و بدون نظارت داده‌کاوی و مقایسه نتایج آنها با یکدیگر برای شناسایی موارد مشکوک به دستکاری کنتور به‌عنوان موضوع مهم در تلفات شبکه آب است.

۲- مواد و روش‌ها

امروزه با توجه به تولید حجم انبوهی از داده‌ها و ذخیره آنها، بازیابی اطلاعات از چنین پایگاه‌های بزرگی، یک نگرانی عمده محسوب می‌شود. خوشبختانه ابزارهای متعددی به‌منظور استخراج اطلاعات مفید از این پایگاه‌های داده و نیاز کاربران ارائه شده است. امکانی که از طریق آن داده‌ها می‌توانند به‌صورت مؤثر ذخیره و استخراج شوند، به‌عنوان داده‌کاوی شناخته می‌شوند (Monika & Amarpreet, 2018).

یادگیری ماشین و روش‌های مورد استفاده در داده‌کاوی را می‌توان به دو گروه با نظارت و بدون نظارت تقسیم کرد. در یادگیری نظارتی، رکوردها به‌صورت برچسب‌های دستکاری و بدون دستکاری (سالم) استفاده می‌شوند. روش‌هایی مانند درخت تصمیم، شبکه عصبی، ماشین‌بردار پشتیبان، نزدیک‌ترین همسایگی و رگرسیون لجستیک در این دسته قرار می‌گیرند. در یادگیری بدون نظارت، مانند روش‌های خوشه‌بندی برچسبی برای مشاهدات در نظر گرفته نمی‌شود. از بین مدل‌های فوق می‌توان درخت تصمیم و رگرسیون لجستیک را با قوانین و فرمول نشان داد و به‌همین دلیل طرز کار آنها ملموس‌تر است در صورتی که سایر روش‌ها را نمی‌توان به‌راحتی فرموله کرد و تنها خروجی قابل‌ملاحظه است و درک چگونگی مدل‌سازی برای خواننده واضح نیست. از طرفی روش‌هایی مانند شبکه عصبی و ماشین‌بردار پشتیبان به‌خوبی با داده‌های غیرخطی و منحنی‌های پیچیده کار می‌کنند. با توجه به گسسته بودن متغیر پاسخ، مدل‌های فوق از پرکاربردترین روش‌ها برای چنین داده‌هایی هستند و در عمل این مدل‌ها برازش می‌شوند و مدل با کمترین خطا انتخاب می‌شود.

۲-۱- درخت تصمیم[1]

درخت تصمیم یک روش داده‌کاوی است که غالباً برای رده‌بندی و پیش‌بینی به‌کار می‌رود. یکی از مزایای درخت تصمیم در رده‌بندی داده‌ها نسبت به سایر روش‌های رده‌بندی مانند شبکه عصبی، سادگی تفسیر و فهم برای تصمیم گیرندگان است. درخت تصمیم در ساختار درختی با شاخه و برگ ارائه می‌شود. ساختار سلسله مراتبی درخت می‌تواند سطوح مختلف فاکتورها را تحلیل کند. هر برگ نشان دهنده نتیجه رده‌بندی و هر شاخه نشانگر شرایط متغیرها است. در ایجاد درخت تصمیم از الگوریتم‌های مختلفی مانند Cart و Chaid استفاده می‌شود. به‌عنوان مثال Cart یک درخت تصمیم دو- دویی است که گوناگونی را به‌عنوان معیار انشعاب قرار می‌دهد و برای این کار از معیار شاخص آنتروپی و یا شاخص جینی استفاده می‌کند. این روش درخت را با کمینه‌سازی خطای تخمینی رده‌بندی نادرست، هرس می‌کند. به‌طور کلی هدف الگوریتم‌های درخت تصمیم، بیشینه‌سازی فاصله بین رده‌هاست. اختلاف این الگوریتم‌ها در معیارهای مختلف فاصله‌ای، روش‌های هرس کردن، وضعیت داده‌های گمشده، تعداد شاخه‌ها در هر گره و نوع داده‌ها بستگی دارد (Hosseini et al., 2013).

۲-۲- شبکه عصبی

یکی از بزرگ‌ترین مزیت‌های شبکه عصبی، انعطاف‌پذیری آنها برای پیش‌بینی انواع مدل‌های غیرخطی است. شبکه عصبی مصنوعی مجموعه‌ای از نرون‌های[2] به‌هم متصل در لایه‌های مختلف است. در ساخت یک مدل بر مبنای شبکه عصبی اولین کار انتخاب نوع شبکه است. پس از آن پارامترهای ورودی که در خروجی تأثیرگذار هستند، انتخاب می شوند. سپس معماری شبکه یعنی تعداد لایه‌ها و تعداد گره‌ها در هر لایه و چگونگی اتصال آنها، و نوع توابع محرکه مورد استفاده برای نرون‌ها و در نهایت پارامترهای مؤثر در آموزش شبکه تعیین می‌شوند. در شکل ۱ یک مدل از شبکه عصبی

[1] Decision Tree
[2] Neuron

doi.org/10.22093/wwj.2020.209875.2958

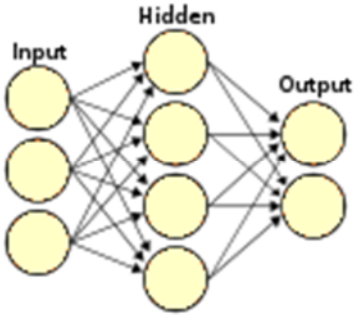

Fig. 1. Neural network with 3 independent input variables, 3 nodes in hidden layer and one output layer

شکل ۱- شبکه عصبی با ۳ متغیر مستقل ورودی، ۴ گره در لایه پنهان و یک لایه خروجی

با ۳ متغیر مستقل ورودی، یک لایه پنهان[1] با ۴ گره[2] و یک لایه خروجی با ۲ گره نمایش داده شده است (Amini and Davood Abadi, 2014).

۲-۳- ماشین‌بردار پشتیبان[3]

ماشین‌بردار پشتیبان یکی دیگر از روش‌های دسته‌بندی داده‌هاست که بر پایه مفهوم صفحات تصمیم هستند که مرز تصمیم را تعریف می‌کنند. یک صفحه تصمیم، داده‌های با برچسب رده مختلف را از هم تفکیک می‌کند. زمانی که برای تفکیک داده‌ها به ساختارهای پیچیده و غیرخطی صفحه تصمیم نیاز است، داده‌های اصلی با به کارگیری مجموعه‌ای از توابع ریاضی به نام کرنل[4] در فضای جدیدی نگاشت داده می‌شوند. در فضای جدید، داده‌های نگاشت داده شده به‌صورت خطی قابل تفکیک هستند. بنابراین به‌جای ساختن یک منحنی پیچیده جداساز، هدف یافتن یک خط بهینه جداساز است. الگوریتم‌های مبتنی بر ماشین‌های‌بردار پشتیبان، الگوریتم‌هایی هستند که سعی می‌کنند یک حاشیه را بیشینه کنند. این الگوریتم‌ها در فضای دو بعدی، برای پیدا کردن خط جداکننده رده‌ها، از دو خط موازی شروع کرده، این خطوط را در جهت خلاف یکدیگر حرکت می‌دهند تا هر کدام از خطوط به یک مشاهده از یک رده خاص برسد. سپس میان دو خط موازی یک نوار یا حاشیه شکل می‌گیرد. هر چه پهنای این نوار بیشتر باشد به این معناست که الگوریتم توانسته هدف بیشینه‌سازی حاشیه را محقق سازد. در مرکز حاشیه، خط جداکننده رده‌ها یا همان خط مرکزی قرار می‌گیرد. حال از بین خطوطی که رسم می‌شوند، خطی که حاشیه کناری آن بیشترین باشد، به عنوان خط جداکننده رده‌ها انتخاب می‌شود. اگر متغیر پاسخ دو مقداری باشد و بردارهای ویژگی کلاس اول روی ابرصفحه H^+ بردارهای ویژگی کلاس دوم روی ابر صفحه H^- قرار گیرند، آنگاه ابر صفحه‌ها به‌صورت معادله ۱ تعریف می‌شوند که در آن بردار وزن w، بردار عمود بر ابرصفحه جداکننده است و b مقدار اریبی است

$$\begin{aligned} H^+ &: w.x + b = +1 \\ H^- &: w.x + b = -1 \end{aligned} \qquad (۱)$$

در واقع معادله یک رده بند ابرصفحه با ناحیه حاشیه بهینه به‌صورت معادله ۲ خواهد بود

$$\begin{aligned} &\text{Min} && \varphi(w) = \frac{1}{2}\|w\|^2 \\ &\text{Subject} && \text{to} \quad y_i(w.x_i + b) \geq 1 \\ &i = 1,2,\ldots,n \end{aligned} \qquad (۲)$$

مسئله فوق، یک مسئله بهینه‌سازی از نوع محدب و درجه دوم است و برای حل آن باید با استفاده از تابع لاگرانژ (معادله ۳)، ضرایب لاگرانژ را به‌دست آورد

$$L(w,b,a) = \frac{1}{2} w.w - \sum_{i=1}^{n} a_i (y_i (w.x_i + b) - 1) \qquad (۳)$$

هر یک از ضرایب لاگرانژ به‌دست آمده متناظر با یکی از الگوهای X است. الگوهای X را که متناظر با ضرایب مثبت آلفا هستند، بردار پشتیبان می‌نامند. همچنین می‌توان با به‌دست آوردن مقدار w و قرار دادن آن در معادله ۳، مسئله را به مسئله دوگان تبدیل کرد و جواب‌های بهینه را به‌دست آورد (Hosseini et al., 2013).

۲-۴- K نزدیک‌ترین همسایه[5]

این روش یک الگوریتم ساده مبتنی بر نمونه است که تمامی

[1] Hidden Layer
[2] Node
[3] Support Vector Machine (SVM)
[4] Kernel Funection
[5] K- Nearest Neighborhood

dx.doi.org/10.22093/wwj.2020.209875.2958

نمونه‌های آموزشی را ترسیم و نمونه‌های بدون برچسب را بر اساس نزدیک‌ترین همسایگی آنها طبقه‌بندی می‌کند. در این روش برای یک داده آزمایشی الگوریتم به دنبال K نمونه مشابه می‌شود. نزدیکی دو نمونه با به‌دست آوردن تشابه و یا فاصله میان این دو نمونه محاسبه می‌شود. پس از یافتن این K داده مشابه با نمونه آزمایشی، با رأی اکثریت برچسب کلاس داده آزمایشی انتخاب می‌شود. در این روش با تغییر K عملکرد طبقه‌بندی نیز تغییر می‌کند. معمولاً پارامتر K به‌صورت تجربی انتخاب می‌شود. به‌عبارتی تعداد مختلفی از نزدیک‌ترین همسایگی‌ها آزمایش شده و پارامتر با بهترین دقت عملکرد برای تعریف طبقه‌بندی انتخاب می‌شود (Hassanat et al., 2014).

۲-۵- رگرسیون لجستیک[1]

از این روش برای مقدار متغیر پاسخ اسمی استفاده می‌شود. رگرسیون لجستیک احتمال وقایع را به‌صورت تابع خطی از متغیرها بیان می‌کند. در حقیقت این مدل به‌جای تخمین مقدار پاسخ، احتمال آن را پیش‌بینی می‌کند. رگرسیون لجستیک با توجه به سطوح متغیر پاسخ، به انواع دو مقداری، چندگانه و ترتیبی می‌شود. تفاوت رگرسیون لجستیک با رگرسیون معمولی در شکل ۲ نشان داده شده است. رگرسیون لجستیک منحنی شکل و رگرسیون خطی به‌صورت یک خط مستقیم است.

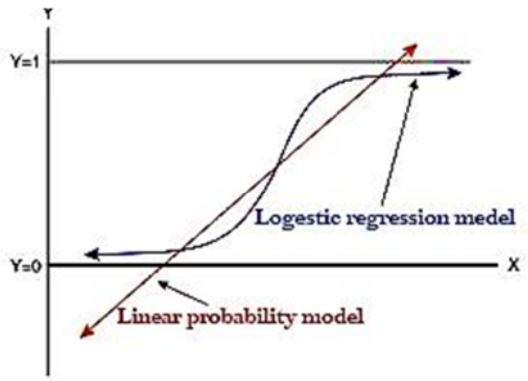

Fig. 2. Logistic curve and linear regression

شکل ۲- منحنی لجستیک و رگرسیون خطی

اگر $x_1, x_2, ..., x_k$ نشان دهنده متغیرهای مستقل باشند و $0 < p < 1$ احتمال تعلق به دسته پاسخ، آنگاه مدل رگرسیون لجستیک با تابع کرنل نقطه‌ای به‌صورت معادله ۴ نشان داده می‌شود (Navanshu et al., 2018)

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + ... + b_k x_k)}} \quad (۴)$$

۲-۶- خوشه‌بندی

استخراج و شناسایی الگو از بین حجم انبوه داده‌ها یک مسئله غیر نظارتی است که با استفاده از روش‌های خوشه‌بندی قابل‌اجرا است. در این راستا راه‌کارهای متعددی ارائه شده است که با استفاده از الگوریتم‌های خوشه‌بندی، داده‌های مشابه را در یک خوشه قرار داده و الگویی را به‌عنوان نماینده ارائه می‌کند. در واقع این نماینده بیانگر رفتار داده‌های آن خوشه است (Amoozagar, 2016). خوشه‌بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه شده است. قابلیت خوشه‌بندی در ورود به فضای داده و تشخیص ساختار آنها، خوشه‌بندی را یکی از ایده‌آل‌ترین مکانیسم‌ها برای کار با دنیای عظیم داده‌ها کرده است (Rastgar, 2010).

در بحث خوشه‌بندی، هدف این است که داده‌هایی که در یک خوشه قرار دارند بیشترین تشابه را با یکدیگر و کمترین تشابه را با اعضای خوشه‌های دیگر داشته باشند. با این مبنای عملکردی، الگوریتم‌های متنوعی برای خوشه‌بندی وجود دارد مانند خوشه‌بندی سلسله مراتبی، K میانگین[2] و خوشه‌بندی دو مرحله‌ای[3]. در میان روش‌های موجود برای خوشه‌بندی داده‌ها، روش‌های K میانگین و سلسله مراتبی (دو مرحله‌ای) دارای شباهت عملکرد بیشتری هستند، ولی با توجه به قابلیت بالای خوشه‌بندی دو مرحله‌ای در استفاده هم‌زمان از متغیرهای کمّی و کیفی، انتخاب بهینه تعداد خوشه‌ها و پوشش خوب داده‌های بزرگ، می‌تواند گزینش بهتری باشد (Kajori et al., 2015).

۲-۷- داده‌های پژوهش

به‌منظور شناسایی کنتورهای دست‌کاری شده، مشترکان به دو گروه دارای سابقه دست‌کاری و بدون سابقه دست‌کاری (سالم) تقسیم شدند. تعداد ۶۷۱ مشترک خانگی که توسط واحد انشعابات غیرمجاز شرکت آب و فاضلاب شهری قم در سال‌های ۹۶ تا ۸ ماه

[1] Logestic Regression

[2] K-means

[3] Two step

doi.org/10.22093/wwj.2020.209875.2958

اول ۹۸ به‌عنوان کنتور دست‌کاری شده شناسایی شدند، در گروه اول قرار گرفتند. نمونه تصادفی ۳۱۲۰ تایی از مشترکان بدون سابقه دست‌کاری کنتور (سالم) نیز به گروه دوم اختصاص یافتند.

متغیر پاسخ دو مقداری با برچسب‌های دست‌کاری کنتور و غیردست‌کاری (سالم) است. متغیرهای مستقل عبارتند از انحراف‌معیار و ضریب تغییرات میانگین مصرف ماهانه، ضریب تغییرات میانگین درآمد، ضریب تغییرات میانگین مبالغ وصول شده و انحراف‌معیار نرخ آب بها در سه سال آخر داده‌های جمع‌آوری شده.

با توجه به این که مصرف رابطه مستقیمی با مدت و تعداد آحاد مشترک دارد، بنابراین به‌منظور همسان‌سازی از میانگین مصرف ماهانه قبوض قطعی به ازای هر واحد خانگی استفاده شد. مشترکانی که در طی سال‌های مورد مطالعه تغییر کاربری و یا تغییر واحد داشتند نیز از پژوهش خارج شدند. برای تحلیل داده‌ها از نرم‌افزار داده‌کاوی Rapidminer استفاده شد.

۳- نتایج و بحث

۳-۱- مدل‌سازی و شناسایی مصارف غیرمجاز آب خانگی

به‌منظور مدل‌سازی و شناسایی کنتورهای دست‌کاری شده، از روش‌های نظارتی مانند درخت تصمیم، ماشین‌بردار پشتیبان، شبکه عصبی، رگرسیون لجستیک و نزدیک‌ترین همسایگی و همچنین روش بدون نظارتی خوشه‌بندی دو مرحله‌ای به‌منظور داده‌کاوی استفاده شد. جدول ۱ آمار توصیفی متغیرهای منتخب را نشان می‌دهد. با توجه به جدول در تمامی متغیرها، مقدار آمارهای توصیفی برچسب دست‌کاری کنتور بزرگتر از برچسب سالم است.

در ادامه به‌منظور استفاده از روش‌های داده‌کاوی، ۸۰ درصد داده‌ها (۳۰۳۳ مورد) برای آموزش و ۲۰ درصد باقی‌مانده (۷۵۸ مورد) برای آزمایش مدل‌ها در نظر گرفته شد. جدول ۲ نتایج استفاده از مدل آموزش دیده بر روی داده‌های آزمایش را نشان می‌دهد. پارامترها با استفاده از روش‌های بهینه‌سازی به‌دست آمدند. در این روش‌ها مقدار پارامترها تا رسیدن به کمترین خطا، تغییر می‌کنند.

با توجه به جدول ملاحظه می‌شود روش رگرسیون لجستیک با دقت کلی ۸۹/۴۵ درصد، بیشترین دقت را در بین سایر روش‌ها دارد. در این روش ۸۵ درصد کنتورهای دست‌کاری شده و ۹۱ درصد کنتورهای فاقد دست‌کاری (سالم) توسط مدل به‌درستی تشخیص داده شده‌اند. درخت تصمیم و نزدیک‌ترین همسایگی بعد از رگرسیون لجستیک دقت کلی ۸۸ درصد دارند ولی درصد پاسخ

جدول ۱- آمار توصیفی متغیرها

Table 1. Descriptive statistics of variables

Variable	Statistics	Response label	
		Normal	Tampering
The standard deviation of mean consumption	Mean	2.06	6.74
	Median	1.45	4.72
	Sd	2.02	7.32
The coefficient of variation of mean consumption	Mean	0.15	0.39
	Median	0.10	0.31
	Sd	0.17	0.31
Coefficient of variation of mean bills paid	Mean	0.36	0.87
	Median	0.28	0.78
	Sd	0.29	0.49
Coefficient of variation of bills amount	Mean	0.24	0.64
	Median	0.19	0.53
	Sd	0.20	0.41
The standard deviation of water rates	Mean	625.94	2805.90
	Median	348.50	870.16
	Sd	1237.08	4841.48

dx.doi.org/10.22093/wwj.2020.209875.2958

جدول ۲- نتایج مدل‌سازی در نمونه تست

Table 2. Modeling results in test sample

Model	Optimal parameters	Observed label	Predicted label		Correct response percentage	Model accuracy (%)
			Tampering	Normal		
Decision Tree J-48	-	Tampering normal	93 37	52 576	64 94	88.26
Support Vector Machine	C= .056 Gamma=.4	Tampering normal	123 70	22 543	85 89	87.86
Neural Network	Learning rate=.7 Momentum=.7	Tampering normal	82 37	63 576	57 94	86.81
Logistic Regression	Gamma= .9 Degree=2 C=.25 Kernel= anova	Tampering normal	123 58	22 555	85 91	89.45
Nearest Neighborhood	K=7	Tampering normal	73 14	72 599	50 98	88.65

جدول ۳- نتایج خوشه‌بندی

Table 3. Clustering results

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Count	1179	519	160	227	337	1369
Coefficient of variation of bills amount	0.18	0.48	1.21	0.24	0.59	0.20
Coefficient of variation of mean consumption	0.09	0.29	0.65	0.16	0.51	0.12
Coefficient of variation of bills paid mean	0.25	0.78	1.10	0.39	0.93	0.31
The standard deviation of mean consumption	1.44	4.42	15.39	1.86	5.59	1.61
The standard deviation of water rates	404.19	1233.66	9576.46	420.98	1455.44	438.71

صحیح به موارد دستکاری آنها کم است. شبکه عصبی نیز با اینکه درصد زیادی از کنتورهای سالم را تشخیص داده ولی در خصوص کنتورهای دستکاری شده مانند درخت تصمیم و نزدیک‌ترین همسایگی ضعیف عمل کرده است. ماشین‌بردار پشتیبان مانند روش رگرسیون لجستیک توانسته موارد دستکاری شده را به‌خوبی شناسایی کند و در موارد کنتورهای سالم کمی ضعیف‌تر از روش رگرسیون لجستیک عمل کرده است. بنابراین برای شناسایی موارد دستکاری شده می‌توان از نتایج روش رگرسیون لجستیک با اطمینان قابل‌توجهی استفاده کرد.

نتایج بالا مربوط به روش‌های با نظارتی است که در آن متغیر پاسخ مشخص است. در ادامه از روش بدون نظارتی خوشه‌بندی دو مرحله‌ای برای دسته‌بندی مشترکان مشابه بدون در نظر گرفتن متغیر پاسخ استفاده شد. متغیرهای مورد استفاده در این مرحله، علاوه بر متغیرهای مستقل استفاده شده در روش‌های قبل، متغیر اسمی مناطق شهری از نظر شرکت به‌دلیل تفاوت الگوی مصرف مناطق، نیز حضور داشتند (Amini et al., 2018b). بعد از اجرای خوشه‌بندی روی تمام داده‌ها، تعداد ۶ خوشه بهینه با کیفیت خوب با استفاده از معیارهای فاصله درست‌نمایی بیشینه و بیزین به‌صورت خودکار تعیین شد. نتایج خوشه‌بندی در جدول ۳ نشان داده شده است.

اعداد جدول نشانگر میانگین متغیرها در هر خوشه است که به‌عنوان نماینده آن خوشه در نظر گرفته می‌شود. خوشه شماره ۳ با تعداد ۱۶۰ اشتراک و مقدار زیاد در هر یک از متغیرها خود را کاملاً از دیگر خوشه‌ها مجزا کرده است. جدول ۴ توزیع فراوانی متغیر پاسخ را در هریک از خوشه‌ها نشان می‌دهد. تعداد ۱۳۷ مورد در خوشه سه، دارای برچسب دستکاری بودند که بیشترین درصد

dx.doi.org/10.22093/wwj.2020.209875.2958

جدول ۴- توزیع فراوانی متغیر پاسخ در خوشه‌ها

Table 4. Frequency distribution of responses in clusters

Label	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Total
Normal	1105	242	23	215	227	1308	3120
Tampering	74	277	137	12	110	61	671
Total	1179	519	160	227	337	1369	3791

دست‌کاری در بین خوشه‌ها است. به‌عبارتی ۸۶ درصد اعضای خوشه ۳، مشترکانی هستند که سابقه دست‌کاری کنتور داشته‌اند. این خوشه ۲۰ درصد کل دست‌کاری‌ها را در خود قرار داده است. بنابراین می‌توان با استفاده از نتایج خوشه‌بندی، لیستی از اشتراک‌ها را تعیین و در پیمایش‌ها به‌عنوان کاندیدی برای بررسی موارد مشکوک به دست‌کاری کنتور استفاده کرد.

جدول ۵ توزیع فراوانی مقایسه بین روش رگرسیون لجستیک به‌عنوان بهترین مدل در روش‌های با نظارت و خوشه‌بندی به‌عنوان روش بدون نظارت را نشان می‌دهد. ۱۸ درصد از موارد دست‌کاری شناسایی شده توسط رگرسیون لجستیک در خوشه شماره ۳ قرار دارد. بنابراین در مواردی که نیاز سریع به پیمایش وجود دارد می‌توان از نتایج حاصل از روش خوشه‌بندی استفاده کرد.

جدول ۵- مقایسه توزیع فراوانی روش رگرسیون لجستیک و خوشه‌بندی در کل داده‌ها

Table 5. Comparison of frequency distribution of logistic regression method and clustering in total data

	Forecast logistic regression	
Cluster	Tampering	Normal
Cluster 1	24	1155
Cluster 2	327	192
Cluster 3	160	0
Cluster 4	41	186
Cluster 5	264	73
Cluster 6	60	1309

۴- نتیجه‌گیری

در این پژوهش با استفاده از تکنیک‌های داده‌کاوی، به مدل‌سازی و شناسایی کنتورهای دست‌کاری شده به‌عنوان مصداقی از مصارف غیرمجاز آب در مشترکان خانگی شهر قم پرداخته شد. متغیرهای استفاده شده، شامل ضریب تغییرات متغیرهای میانگین مصرف، میانگین درآمد، میانگین وصول و همچنین انحراف معیار متغیرهای میانگین مصرف و نرخ آب‌بها بودند. مقایسه روش‌های مختلف داده‌کاوی مانند درخت تصمیم، شبکه عصبی، ماشین‌بردار پشتیبان، رگرسیون لجستیک و نزدیک‌ترین همسایگی نشان داد رگرسیون لجستیک با دقت ۸۹ درصد بهترین روش برای مدل‌سازی است. این روش توانست ۸۵ درصد موارد دست‌کاری شده و ۹۱ درصد موارد غیر دست‌کاری (سالم) را به درستی تشخیص دهد.

با استفاده از روش خوشه‌بندی دو مرحله‌ای نیز به‌عنوان یک روش بدون نظارت و بدون نیاز به متغیر پاسخ، ۶ خوشه تعیین شد. بر اساس متغیرهای ورودی برای خوشه‌بندی، خوشه شماره ۳ با تعداد ۱۶۰ اشتراک، رفتاری مجزا از سایر خوشه‌ها از خود نشان داد. ۸۶ درصد اعضای این خوشه، مشترکانی هستند که سابقه دست‌کاری کنتور داشته‌اند. این خوشه ۲۰ درصد کل دست‌کاری‌ها را در خود قرار داده است.

روش‌های با نظارت مانند رگرسیون لجستیک به‌دلیل وجود متغیر پاسخ دقت و کارایی زیادی نسبت به روش‌های بدون نظارت مانند خوشه‌بندی دارند. اما در جایی که به نحوی دسترسی به سابقه کنتورهای دست‌کاری شده امکان‌پذیر نباشد و یا نیاز سریع به بررسی و پیمایش مشترکان وجود داشته باشد، با استفاده از روش خوشه‌بندی می‌توان مواردی را شناسایی و به‌عنوان کاندیدی برای موارد دست‌کاری کنتور معرفی کرد.

۵- قدردانی

نویسنده مقاله از حمایت‌های معاونت محترم مشترکان و دفتر آمار و تحلیل اطلاعات مشترکان تشکر می‌کند.

References

Amini, G. & Davood Abadi, A. 2014. Estimating household water demand of the city of Qom using artificial neural networks and log linear regression. *1st Water Sciences and Engineering Conference, Tehran, Iran*. (In Persian).

Amini, G., Entezam, H., Sadeghpour, A. & Davood Abadi, A. 2018a. Application of data mining to identify subscribers with unauthorized use of water (case study of Qom water and wastewater company). *2nd Iran Water and Wastewater Science Engineering Congress and National Conference on Demand & Supply of Drinking Water and Sanitation*, Isfahan, Iran. (In Persian).

Amini, G., Entezam, H., Sadeghpour, A. & Davood Abadi, A. 2018b. Identification and extraction of water consumption patterns by data mining (Case study of Qom water and wastewater company). *2nd Iran Water and Wastewater Science Engineering Congress and National Conference on Demand & Supply of Drinking Water and Sanitation*, Isfahan, Iran. (In persian).

Amini, G. & Saeidi, Z. 2017. Identification of meteorological parameters affecting water consumption in household sector of Qom. *Journal of Water and Wastewater*, 29(2)**,** 48-58. (In Persian)

Amoozagar, M. 2016. Provides a two-step solution to identify the pattern of power consumption. *Iranian Electric Industry Journal of Quality and Productivity*, 5**,** 48-57. (In persian).

Anita, B. D. & Ravindra, D. 2013. Data mining techniques for fraud detection. *Journal of Computer Science and Information Technologies*, 4**,** 1-4.

Hashem, E. & Humaid, S. 2012. A data mining based fraud detection model for water consumption billing system in MOG. MSc Thesis, Islamic University of Gaza.

Hassanat, A. B., Abbadi, M. A., A., A. G. & Alhasanat, A. A. 2014. Solving the problem of the k parameter in the knn classifier using an ensemble learning approach. *Journal of Computer Science and Information Security*, 12**,** 33-39.

Hosseini, R., Sarmad, M. & Noghabi, M. 2013. Data mining in r by rattle package. *Jornal of Andishe- ye Amari*, 35**,** 17-29. (In Persian)

Kajori, M., Feriedunian, A. & Lesani, H. 2015. Identifying the pattern of electric energy consumption with data mining. *30th International Electrical Conference***,** Tehran, Iran. (In Persian).

Kasaeyan, A. & Ghayni, M. 2017. Examining unauthorized consumption detection methods based on measurement data in intelligent network structure. *32nd International Electrical Conference***,** Tehran, Iran. (In Persian).

Minaie, B., Dianat, R., Hani, H. & Sobhaninia, M. 2011. Identify fraudsters in service organizations using data mining. MSc Information Technology, University of Qom, Iran. (In Persian).

Monedero, I., Biscarri, F., Guerrero, J., Roldan, M. & Leon, C. 2015. An approach to detection of tampering in water meters. *19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems*, National Uiversity of Singapore, Singapore.

Monika, C. & Amarpreet, K. 2018. A comparative study of classification techniques for fraud detection. *Journal on Future Revolution in Computer Science & Communication Engineering*, 4**,** 19-23.

Navanshu, K. & Saad, Y. S. 2018. Credit card fraud detection using machine learning modeles and collating machine learning models. *Journal of Pure and Applied Mathematics*, 118**,** 825-838.

Rastgar, H. 2010. Investigating aggregation clustering algorithms and simulating and executing a sample. MSc Thesis, Payame Noor University of Mashhad, Mashhad, Iran. (In Persian)