

پیش‌بینی شکست لوله‌ها با مدل رگرسیون لجستیک وزن‌دار برای مدیریت داده‌های نامتعادل و رویدادهای مکرر (مطالعه موردی: منطقه ۲ شهری تهران)

محسن ستاری^۱، همایون مطیعی^{۲*}

۱- دانشجوی دکتری، دانشکده عمران، آب و محیط زیست، دانشگاه شهید بهشتی، تهران، ایران.

(m.sattari@sutech.ac.ir)

۲- دانشیار، دانشکده عمران، آب و محیط زیست، دانشگاه شهید بهشتی، تهران، ایران.

(h_motiei@sbu.ac.ir)

* نویسنده مسئول

تاریخ دریافت: ۱۴۰۴/۰۸/۲۹

تاریخ اصلاح: ۱۴۰۵/۰۳/۱۸

تاریخ پذیرش: ۱۴۰۵/۰۴/۰۱

چکیده

پیش‌بینی شکست لوله‌ها در شبکه‌های توزیع آب، برای گذار از نگهداری واکنشی به مدیریت پیشگیرانه، ضروری است. اما توسعه مدل‌های پیش‌بینی دقیق با چالش‌های داده‌های نامتعادل (شکست و عدم شکست) و ماهیت تکرارشونده آنها مواجه است که اغلب نادیده گرفته می‌شوند. این پژوهش با توسعه یک مدل رگرسیون لجستیک وزن‌دار به عنوان یک تکنیک یادگیری ماشین، به این چالش‌ها پاسخ می‌دهد. در تحقیق حاضر از داده‌های واقعی شکست لوله‌ها در شبکه توزیع آب منطقه ۲ شهری تهران به مدت ۱۲ سال استفاده شد. داده‌ها به صورت تصادفی به دو زیرمجموعه ۷۰ درصد برای آموزش و ۳۰ درصد برای اعتبارسنجی تقسیم شد تا عملکرد مدل بر روی داده‌های دیده‌نشده، صحت سنجی گردد. برای ارزیابی مدل از منحنی مشخصه عملکرد استفاده و مدل با مساحت زیر نمودار برابر ۰.۸۸ در مجموعه اعتبارسنجی، قدرت پیش‌بینی بالایی را از خود نشان داد. طول لوله با نسبت شانس ۱/۰۲ مهم‌ترین پیش‌بینی‌کننده شناسایی شد. ضریب منفی سن لوله (۰/۱۰۴-) یک یافته کلیدی بود که به سوگیری بازماندگان ناشی از شکست‌های زودهنگام به دلیل نصب نامناسب، نسبت داده شد. این مدل می‌تواند ابزاری قابل اعتماد برای شرکت‌های آب جهت اولویت‌بندی بازرسی و جایگزینی لوله‌ها بوده و بر نقش حیاتی کیفیت نصب در کنار ویژگی‌های فیزیکی لوله، تأکید دارد. کلمات کلیدی: شکست لوله، شبکه توزیع آب، یادگیری ماشین، داده‌های نامتعادل، رویدادهای تکرارشونده.

Pipe Failure Prediction Using Weighted Logistic Regression for Managing Imbalanced and Recurring Events (Case Study: District 2 of Tehran)

Mohsen Sattari¹, Homayoun Motiee^{2*}

1- Ph.D. Candidate, Civil, Water and Environmental faculty, Shahid Beheshti University, Tehran, Iran.

(m.sattari@sutech.ac.ir)

2-Associate Professor, Civil, Water and Environmental faculty, Shahid Beheshti University, Tehran, Iran

(h_motiei@sbu.ac.ir)

*Corresponding author

Received: 20/11/2025

Revised: 08/06/2026

Accepted: 22/06/2026

Abstract

Predicting pipe failures in water distribution networks is crucial for transitioning from reactive maintenance to proactive management. However, developing accurate predictive models faces challenges posed by imbalanced data (failures vs. non-failures) and the recurring nature of these events, which are often overlooked. This research addresses these challenges by developing a weighted logistic regression model, a machine learning technique. The study utilized 12 years of real-world pipe failure data from the water distribution network of District 2 in Tehran. The data were randomly split into two subsets: 70% for training and 30% for validation, to verify the model's performance on unseen data. The model's performance was evaluated using the Receiver Operating Characteristic (ROC) curve, and it demonstrated high predictive power with an Area Under the Curve (AUC) of 0.88 on the validation set. Pipe length was identified as the most significant predictor, with an odds ratio of 1.02. A negative coefficient for pipe age (-0.104) was a key finding, attributed to survivor bias stemming from premature failures due to improper installation. This model can serve as a reliable tool for water utilities to prioritize pipe inspections and replacements, emphasizing the critical role of installation quality alongside the pipe's physical characteristics.

Keywords: Pipe Failure, Water Distribution Network, Machine Learning, Imbalanced Data, Recurrent Events.

۱- مقدمه

شبکه‌های توزیع آب از مهم‌ترین اجزای زیرساخت‌های شهری مدرن به شمار می‌آیند و نقشی کلیدی در تأمین مستمر آب آشامیدنی برای بخش‌های مختلف از جمله مصارف خانگی، صنعتی و کشاورزی دارند. این سامانه‌های وسیع که عمدتاً در زیر سطح زمین قرار گرفته‌اند، نشان‌دهنده سرمایه‌گذاری‌های قابل توجه و بلندمدت در حوزه زیرساختی هستند. با این حال، بسیاری از این شبکه‌ها به دلیل عبور از دوره بهره‌وری مطلوب و قرار گرفتن در مرحله فرسودگی، بیش از پیش در معرض آسیب، خرابی و وقوع اختلالات ناشی از نقص‌های فنی قرار دارند (Lešťáková et al. 2024).

افت فشار، نارضایتی مشترکین، تلفات سنگین اقتصادی، مدیریت پرهزینه و فرسایش بسترهای لوله را می‌توان از جمله اثرات شکست لوله‌ها در نظر گرفت که به منظور جلوگیری از اتلاف انرژی و کمک به حفظ تامین آب باید کاهش شکست را به عنوان اولویت اول و عاملی جهت افزایش پایداری آب شهرها در نظر گرفت (نوذری و همکاران، ۱۴۰۲). صنعت آب در حال حاضر در مسیر تغییر از رویکردهای سنتی نگهداری واکنشی به سمت راه‌حل‌های پیش‌بینی‌کننده‌تر و پیشگیرانه‌تر روی آورده است. نگهداری واکنشی شامل نظارت، تعمیر و تعویض لوله‌ها در پاسخ به خرابی‌ها می‌شود. با این حال، یک رویکرد پیشگیرانه‌تر شامل مدل‌سازی زوال شبکه و ارزیابی خطر خرابی در اجزای آن برای جلوگیری از ترکیدگی است. از طریق استفاده از روش‌های داده‌محور و تجزیه و تحلیل پیش‌بینی‌کننده، ارزیابی قابلیت اطمینان شبکه‌های توزیع آب و پیش‌بینی وقوع خرابی‌ها امکان‌پذیر می‌شود. در نتیجه، این امر امکان اجرای نگهداری پیش‌بینی‌کننده را به عنوان جزئی از نگهداری پیشگیرانه فراهم می‌کند (Latifi et al. 2024).

ارزیابی دقیق یکپارچگی پیش‌بینی در سیستم‌های خط لوله برای اجرای استراتژی‌های نگهداری پیشگیرانه که از خرابی‌های فاجعه‌بار جلوگیری کرده و قابلیت اطمینان خدمات را تضمین می‌کنند، بسیار مهم است. در دهه‌های اخیر، کاربرد یادگیری ماشین^۱ به عنوان یک تکنیک امیدوارکننده برای پردازش و استخراج تعاملات پیچیده بین عوامل مؤثر و روندهای خرابی در سیستم‌های شبکه توزیع آب شهری ظهور کرده است (Chen et al. 2025). مدل‌های مبتنی بر یادگیری ماشین به دلیل مهارتشان در مدیریت مجموعه داده‌های گسترده و پیچیده، به عنوان ابزاری برجسته در حوزه پیش‌بینی شکست هستند. نقاط قوت اصلی آنها شامل دقت بالا، قابلیت‌های پیش‌بینی قوی، سازگاری و پتانسیل یادگیری و بهبود مداوم بر اساس داده‌های بازخورد است (Wang et al. 2022). چندین مدل پیش‌بینی (فیزیکی و آماری) در طول دهه‌های گذشته برای پیش‌بینی نرخ شکست و تخریب خطوط لوله آب توسعه داده شده‌اند. مدل‌های فیزیکی برای یافتن اجزای مؤثر بر شکست لوله به داده‌های ورودی قابل توجهی نیاز دارند. بنابراین، توسعه این مدل‌ها پرهزینه و زمان‌بر است. از سوی دیگر، مدل‌های آماری، شکست لوله را با استفاده از داده‌های تاریخی پیش‌بینی می‌کنند و در مقایسه با مدل‌های فیزیکی، ساده‌تر و کم‌هزینه‌تر هستند. الگوریتم‌های داده‌کاوی و یادگیری ماشین می‌توانند به طور قابل ملاحظه‌ای به اولویت‌بندی بازرسی‌های میدانی، بازسازی‌ها و تعویض‌ها کمک کرده و از تعداد زیادی از شکستگی‌های لوله جلوگیری کنند (Snider and McBean, 2020).

در سال‌های اخیر، محققان از الگوریتم‌های یادگیری ماشین برای پیش‌بینی نرخ شکستگی خطوط لوله آب استفاده کرده‌اند. روش‌های رایج مورد استفاده شامل تحلیل رگرسیون^۲، درخت تصمیم^۳ (DT)، جنگل تصادفی^۴ (RF)، ماشین‌های بردار پشتیبان^۵ (SVM) و تقویت گرادینان شدید^۶ (XGboost) هستند (Barton et al, 2022).

در تحقیق (Omar et al, 2023) شش مدل یادگیری ماشین برای پیش‌بینی خرابی‌ها در شبکه توزیع آب شهر کیچنر کانادا مقایسه شد. نتایج مطالعه نشان داد که پایش و مداخله‌ی پیشگیرانه در تنها ۸ درصد از کل سامانه می‌تواند حدود ۷۲ درصد از خرابی‌ها را کاهش دهد.

Robles-Velasco et al. (2020) از رگرسیون لجستیک^۷ (LR) و طبقه‌بندی بردار پشتیبان^۸ (SVC) برای پیش‌بینی احتمال خرابی استفاده کردند. آنها از یک تکنیک نمونه‌گیری کمتر برای مقابله با مجموعه داده‌های نامتوازن استفاده کردند. علاوه بر این، آنها مدل‌ها را با توجه به نوع ماده همگن کردند. مطالعه آنها عملکرد کمی بهتر برای LR در مقایسه با SVC را نشان داد و همچنین، برای مطالعه موردی خود، آنها دریافتند که با جایگزینی تنها ۳٪ از خطوط لوله، می‌توان از حدود ۳۰٪ از خرابی‌ها جلوگیری کرد.

(Karadirek et al. 2024) با بهره‌گیری از رگرسیون لجستیک، خرابی لوله‌ها را بر اساس داده‌های دو ساله‌ی رخدادهای ثبت شده پیش‌بینی کردند. در این مدل، عواملی نظیر عمر لوله، قطر، جنس و فشار عملیاتی مورد توجه قرار گرفت. مدل حاصل دقتی برابر با ۷۰،۱٪ به دست آورد و نشان داد که لوله‌های قدیمی‌تر و همچنین آن‌هایی که از مواد PVC یا HDPE ساخته شده‌اند، به‌طور معناداری بیشتر مستعد خرابی هستند، در حالی که لوله‌های چدنی انعطاف‌پذیر از مقاومت بالاتری برخوردارند.

Karimian et al. (2021) یک روش رگرسیون چندجمله‌ای تکاملی برای پیش‌بینی تعداد شکستگی‌های خط لوله آب توسعه دادند. متغیرهای مستقل از طریق بهترین رگرسیون زیرمجموعه انتخاب و خطوط لوله بر اساس طول، سن، جنس و قطر آنها خوشه‌بندی شدند. این مطالعه نشان داد که خطوط لوله با قطر کوچکتر احتمال خرابی بیشتری دارند. همچنین یک تحلیل حساسیت بین طول، جنس و قطر انجام و نتیجه‌گیری گردید که مورد دوم بیشترین حساسیت را روی تعداد پیش‌بینی‌شده‌ی شکستگی‌ها دارد.

برخی از محققین رگرسیون چندجمله‌ای تکاملی را با خوشه‌بندی k-means برای پیش‌بینی شکست خطوط لوله ترکیب کردند و با ادغام این دو روش دقت پیش‌بینی شکست را در مقایسه با رویکرد رگرسیون چندجمله‌ای تکاملی غیرخوشه‌ای به طور قابل توجهی بهبود بخشیدند (Farmani et al. 2017; Kakoudakis et al 2017). این پژوهش‌ها به طور کلی نشان داده‌اند که عواملی مانند طول، قطر، سن و جنس لوله، پیش‌بینی‌کننده‌های کلیدی شکست هستند.

یک چالش مهم در کاربرد مدل مبتنی بر یادگیری ماشین، عدم توجه کافی به اندازه و ویژگی‌های مجموعه داده‌ها در طول انتخاب مدل است. این غفلت اغلب منجر به بروز مشکلاتی در کارایی و کاربردپذیری می‌شود. به عنوان مثال، مدل‌های مبتنی بر درخت و بردار پشتیبان معمولاً در مجموعه داده‌های کوچک تا متوسط عملکرد خوبی دارند و دقت بالایی را با نمونه‌های کمتر ارائه می‌دهند. با این حال، در مجموعه داده‌های بزرگ، این مدل‌ها ممکن است به دلیل تنظیم پیچیده فرامپارامتر و تقاضای محاسباتی بالا، کارایی

کمتری داشته باشند. مدل‌های پیچیده‌تر مانند جنگل تصادفی و درخت تقویت شده ممکن است دقت بالایی ارائه دهند، اما در معرض خطر بیش‌برازش^۹ قرار دارند و تفسیرپذیری کمتری دارند (Solomatine and Ostfeld, 2008). مدل رگرسیون لجستیک به دلایل متعددی برای پیش‌بینی شکست لوله‌ها گزینه‌ی بسیار مناسبی محسوب می‌شود. این روش، قابلیت اطمینان بالایی در انجام وظایف طبقه‌بندی دودویی^{۱۰} دارد؛ به عنوان مثال، می‌تواند به طور دقیق پیش‌بینی کند که آیا یک لوله در معرض شکست قرار خواهد گرفت یا خیر (Rifaai et al, 2022). با توجه به اینکه این مدل قادر به تخمین احتمال وقوع یک رویداد بر اساس ویژگی‌های ورودی است، در درک رابطه بین متغیرهای مختلف و نتیجه‌ی مورد نظر (شکست لوله) بسیار مؤثر عمل می‌کند. علاوه بر این، رگرسیون لجستیک تفسیرپذیری نتایج را فراهم می‌آورد و امکان درک تأثیر هر ویژگی بر احتمال خرابی لوله را میسر می‌سازد. این قابلیت تفسیر برای تصمیم‌گیری آگاهانه در مدیریت زیرساخت‌ها بسیار مهم است (Asadi, 2024). با اینکه این مدل، به راحتی قابل تفسیر هست اما ممکن است در مدل‌سازی روابط غیرخطی موجود در داده‌ها ناکارآمد باشد (Barton et al, 2022).

با وجود پیشرفت‌های قابل توجه در حوزه داده‌کاوی و یادگیری ماشین، پژوهش‌های گذشته در زمینه مدل‌سازی شکست لوله‌های آب با چالش‌های متعددی روبرو بوده‌اند که منجر به یک شکاف تحقیقاتی مشخص بین پژوهش‌های آکادمیک و نیازهای عملی دنیای واقعی شده و این امر اثربخشی مدل‌ها را محدود کرده است. بسیاری از مطالعات پیشین، به طور همزمان با سه چالش کلیدی روبرو هستند که کارایی عملی آن‌ها را محدود می‌کند. چالش اول، عدم تعادل کلاس‌ها است؛ در داده‌های شکست، رویداد شکست (کلاس اقلیت) بسیار نادرتر از عدم شکست (کلاس اکثریت) است که این امر مدل‌ها را به سمت پیش‌بینی کلاس اکثریت سوگیری می‌دهد. در واقع در مجموعه داده‌های واقعی اتفاقات شبکه توزیع آب، تعداد مشاهدات مربوط به لوله‌های سالم به مراتب بیشتر از لوله‌های دچار شکست است. این عدم تعادل باعث می‌شود مدل‌های یادگیری ماشین تمایل به پیش‌بینی کلاس اکثریت (لوله‌های سالم) پیدا کرده و در شناسایی کلاس اقلیت (لوله‌های شکسته) عملکرد ضعیفی از خود نشان دهند. چالش دوم، نادیده گرفتن ماهیت تکرارشونده شکست‌ها است؛ یک لوله واحد ممکن است چندین بار شکست بخورد که این خود یک سیگنال قوی از ضعف در آن لوله است. در عمل وقتی یک لوله دچار شکست شده، ممکن است به دلایل ساختاری یا محیطی، در معرض ریسک بالاتری برای شکست‌های آتی قرار داشته باشد. نادیده گرفتن این ویژگی تکرارشونده، منجر به از دست رفتن اطلاعات ارزشمندی در مورد پویایی خرابی لوله‌ها و کاهش دقت پیش‌بینی می‌شود. چالش سوم، ضعف در فرآیند صحت‌سنجی است؛ بسیاری از مطالعات، عملکرد مدل خود را صرفاً بر روی داده‌های آموزشی گزارش می‌دهند که این امر، پتانسیل بیش‌برازش را افزایش می‌دهد و در مورد توانایی مدل در تعمیم نتایج به داده‌های جدید، تردید جدی ایجاد می‌کند.

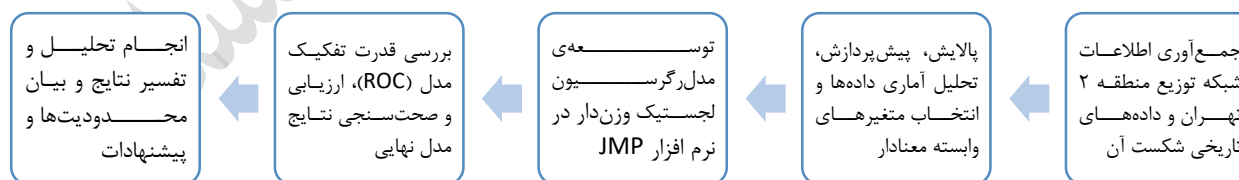
این پژوهش تلاش دارد تا به طور مستقیم به این شکاف تحقیقاتی سه‌گانه پاسخ دهد. هدف اصلی، توسعه یک چارچوب مدل‌سازی

جامع و قابل اعتماد برای پیش‌بینی شکست لوله‌های آب است که همزمان این سه چالش را هدف قرار دهد. در حوزه مدیریت زیرساخت‌های حیاتی مانند شبکه‌های آب، صرفاً ارائه یک پیش‌بینی کافی نیست؛ بلکه درک چرایی و چگونگی رسیدن به آن پیش‌بینی برای تصمیم‌گیرندگان، مهندسان و مدیران اهمیت حیاتی دارد. در این مطالعه، با استفاده از یک مجموعه داده واقعی از شبکه توزیع آب تهران برای مدت ۱۲ سال (بین سال‌های ۱۳۹۲ تا ۱۴۰۳)، یک مدل رگرسیون لجستیک به عنوان یک تکنیک یادگیری ماشین تفسیرپذیر، توسعه یافته است. نوآوری اصلی این مقاله در تلفیق سه رویکرد تلفیقی برای بهبود پیش‌بینی شکست لوله‌های آب شهری است: نخست، استفاده از تکنیک وزن‌دهی مجدد مبتنی بر یادگیری حساس به هزینه، که با اختصاص وزن به نمونه‌های شکست، مشکل عدم تعادل کلاس‌ها را به طور مؤثر حل می‌کند؛ دوم، تغییر ساختار داده‌ها به قالب رویداد-زمان، که امکان در نظر گرفتن شکست‌های تکرار شونده و رفتار دینامیک لوله‌ها را فراهم می‌آورد و در نهایت، بهره‌گیری از مدل تفسیرپذیر رگرسیون لجستیک در کنار فرآیند صحت‌سنجی قوی برای ارزیابی دقیق عملکرد و قابلیت تفسیر نتایج، که در حوزه مدیریت زیر ساخت‌های آب اهمیت زیادی دارد. این ترکیب منحصر به فرد، چارچوبی جامع و قابل اعتماد برای پیش‌بینی و تحلیل شکست لوله‌ها ارائه می‌دهد که نه تنها در بهبود دقت پیش‌بینی، بلکه در ارتقاء قابلیت تصمیم‌گیری‌های عملیاتی و سیاست‌گذاری‌های نگهداری، کاربردی و کارآمد است.

۲- مواد و روش‌ها

این پژوهش با هدف توسعه یک مدل پیش‌بینی‌کننده برای شکست لوله‌های آب، با رویکردی داده-محور و با استفاده از مدل رگرسیون لجستیک انجام شده است. فرآیند کلی تحقیق شامل جمع‌آوری و پالایش داده‌های تاریخی شکست، پیش‌پردازش و مهندسی ویژگی، توسعه مدل و ارزیابی عملکرد آن بوده است که در ادامه به تفصیل شرح داده می‌شود.

در شکل ۱ روندنمای پژوهش حاضر نشان داده شده است.

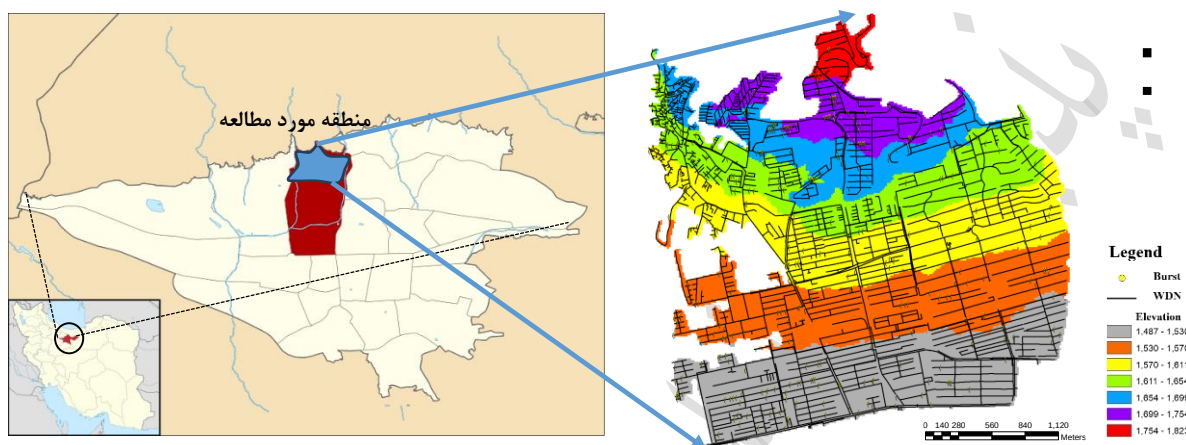


شکل ۱- روندنمای پژوهش حاضر

۲-۱- منطقه مورد مطالعه

جهت بررسی و ارزیابی حوادث رخ داده از آمار حوادث روی داده در شبکه‌ی توزیع آب منطقه ۲ شهری تهران شامل ناحیه ۱

(سعادت آباد و دریا)، ناحیه ۸ (آسمانها و سپهر) و ناحیه ۹ (فراز، پرواز و فرحزاد) که بخشی از منطقه ۱ شرکت آب و فاضلاب تهران می‌باشند، استفاده شده است. منطقه مورد مطالعه، با ۴۷/۱ کیلومترمربع مساحت، در محدوده میانی و شمالی شهر تهران و در ارتفاع حدود ۱۴۵۰-۱۸۵۰ متر از سطح آزاد دریا قرار دارد. منطقه مورد مطالعه از شما به رشته کوه البرز، از شرق به بلوار بیست و چهار متری، از غرب به بزرگراه یادگار امام و از جنوب به بلوار دادمان محدود می‌شود (شکل ۲).



شکل ۲- محدوده مورد بررسی (منطقه ۱ شرکت آب و فاضلاب تهران، منطقه ۲ شهری تهران)

در این تحقیق مجموع طول خطوط لوله این نواحی حدوداً ۲۰۰ کیلومتر می‌باشد که از جنس‌های چدن داکتیل^{۱۱} (DI)، پلی اتیلن^{۱۲} (PE)، آزیست^{۱۳} (AC)، فولادی^{۱۴} (Steel) و یوپی‌وی‌سی^{۱۵} (UPVC) تشکیل شده است. آمار اطلاعات لوله‌ها و حوادث رخ داده (به تعداد مجموع ۵۷۸ حادثه) موضوع این پژوهش در شبکه توزیع آب در منطقه مورد مطالعه برای جنس لوله‌های مختلف در جدول شماره ۱ ارائه شده است.

جدول ۱- مشخصات و تعداد شکست‌ها لوله‌ها بین سال‌های ۱۳۹۲ تا ۱۴۰۳

جنس لوله	تعداد لوله‌ها	مجموع طول (متر)	درصد لوله در شبکه	تعداد شکست	فرکانس شکست در هر متر لوله
AC	۴۱	۱,۶۳۶/۶	۰/۸	۲۱	۰/۰۱۲۸
DI	۲۵۶۹	۱۳۹,۱۵۶	۶۹/۵	۳۱۹	۰/۰۰۲۳
PE	۱۴۱۳	۵۸,۰۵۴	۲۹	۲۳۲	۰/۰۰۴
ST	۱۷	۱,۲۸۲/۴	۰/۶	۳	۰/۰۰۲۳
UPVC	۹	۸۳/۶	۰	۳	۰/۰۳۵۹
مجموع	۴۰۴۹	۲۰۰,۲۱۲	۱۰۰	۵۷۸	---

همانطور که دیده می‌شود، لوله‌های چدن داکتیلی (DI) با ۶۹٪/۵، بیشترین سهم را در ساختار شبکه دارند، در حالی که لوله‌های

آزبست سیمانی (AC) و یوپی‌وی سی (UPVC) با وجود سهم کمتر، بالاترین فرکانس شکست را به ازای هر متر طول لوله به خود اختصاص داده‌اند.

۲-۲- پیش‌پردازش داده‌ها و مهندسی ویژگی

پیش از توسعه مدل، مجموعه داده‌های خام تحت چندین مرحله پیش‌پردازش قرار گرفت تا برای الگوریتم یادگیری ماشین مورد استفاده در این پژوهش (رگرسیون لجستیک) آماده شود. این مراحل عبارت بودند از:

۱. تعریف متغیر وابسته: متغیر هدف، شکست^{۱۶} به صورت یک متغیر دودویی تعریف شد؛ مقدار ۱ برای لوله‌هایی که حداقل یک بار در دوره مطالعه دچار شکست شده‌اند و ۰ برای سایر لوله‌ها. لازم به ذکر است منظور از شکست در این تحقیق شکست فیزیکی لوله‌ها در شبکه توزیع، می‌باشد.

۲. مدیریت رویدادهای تکرار شونده: برای لحاظ کردن اطلاعات مربوط به شکست‌های مکرر یک لوله، ساختار داده‌ها به یک قالب رویداد-زمان^{۱۷} تبدیل شد. در این ساختار، هر شکست ثبت‌شده برای یک لوله به عنوان یک مشاهده مجدد در نظر گرفته شد و سن لوله در زمان وقوع آن شکست به‌روزرسانی گردید.

۳. مدیریت عدم تعادل کلاس‌ها: مجموعه داده به شدت نامتعادل بود (تنها ۱۱٪/۳ لوله‌ها دچار شکست شده‌اند). برای جلوگیری از سوگیری مدل به سمت کلاس اکثریت، از تکنیک وزن‌دهی مجدد^{۱۸} استفاده شد. در این روش، به کلاس اقلیت (لوله‌های شکسته) وزن بیشتری داده می‌شود تا مدل به شناسایی دقیق‌تر آن‌ها ترغیب شده و ضمن حفظ تفسیرپذیری مدل، به طور مؤثری با مشکل عدم تعادل داده‌ها مقابله کند. برای این منظور به هر مشاهده مربوط به کلاس شکست، وزنی معادل با $7/82$ که برابر با نسبت تعداد لوله‌های سالم (۳۵۹۰) به لوله‌های دچار شکست بدون تکرار شکست (۴۵۹) است، اختصاص داده شد. این رویکرد که به عنوان یادگیری حساس به هزینه شناخته می‌شود، هزینه خطای دسته‌بندی اشتباه یک لوله شکسته را افزایش داده و مدل را تشویق می‌کند تا حساسیت بیشتری در شناسایی این کلاس نشان دهد (He and Garcia, 2009).

با توجه به آمار اخذ شده از شرکت آب و فاضلاب منطقه ۱ (اطلاعات و مشخصات شبکه توزیع آب نواحی ۱، ۸ و ۹ منطقه ۲ شهر تهران) متغیرهای مستقل در توسعه مدل شامل؛ طول، عمر، قطر، عمق نصب، فشار و جنس لوله‌ها بوده است.

لازم به ذکر است با توجه به اینکه فشار در شبکه توزیع دارای مقدار ثابتی نیست و علاوه بر این، تأثیر فشار هیدرولیکی بر دقت مدل پیش‌بینی ترکیدگی لوله مبتنی بر داده، حائز اهمیت است (مصلحی و جلیلی، ۱۳۹۵)، در این پژوهش براساس مطالعات قبلی

پژوهشگران در این زمینه، از فشار هیدرولیکی متوسط استفاده شده است (Shirzad and Safari, 2019).

در جدول ۲ نمونه اطلاعات ورودی به نرم افزار برای توسعه مدل، نشان داده شده است.

جدول ۲- نمونه اطلاعات ورودی به نرم افزار برای توسعه مدل

Label	Dep (m)	Dia (mm)	Material	Len (m)	Pr (mH ₂ O)	Age (year)	Weight	Failure	Validation	Failure_Count
P-277	1.1	60	DI	14.3	38.5	20.01644	1	0	0	0
layelole-280	1.1	60	DI	14.5	62.5	27.02192	1	0	1	0
P-128	1.1	60	DI	15.2	36.5	21.03836	1	0	0	0
layelole-2407	1.1	63	UPVC	13	69	29.07123	7.82	1	0	1
layelole-4860	1.3	63	PE	13.1	48.5	21.07397	1	0	1	0
layelole-2475	1.1	63	PE	13.2	70	20.06027	1	0	0	0
layelole-3968	1.2	63	PE	13.8	45	21.05753	1	0	0	0

لازم به ذکر است در داده‌های منطقه مورد مطالعه لوله‌های با طول کمتر از ۲ متر (لوله‌های مربوط به انشعابات که تأثیری در شبکه توزیع ندارند) و همچنین با قطر کمتر از ۶۰ میلیمتر (حداقل قطر در شبکه ۸۰ میلیمتر می‌باشد) از داده‌های ورودی حذف شدند.

در جدول ۳ توصیف آماری لوله‌های موجود شبکه (حداقل، حداکثر و متوسط پارامترها) ارائه شده است.

جدول ۳- توصیف آماری پارامترهای در نظر گرفته شده

پارامتر	طول (متر)	قطر (میلی متر)	سن (سال)	عمق (متر)	فشار (متر آب)
حداکثر	۸۵۱/۴	۷۰۰	۴۵	۲/۲	۹۰
حداقل	۲	۶۰	۰/۰۴	۰/۵	۲
متوسط	۴۹/۶	۱۵۵	۲۳/۸	۱/۲	۵۱،۸

۳-۲- توسعه مدل پیش‌بینی کننده

در این پژوهش، مدل رگرسیون لجستیک به عنوان یک تکنیک یادگیری ماشین تفهیم پذیر و کلاسیک برای پیش‌بینی احتمال شکست لوله به کار گرفته شد. این مدل به دلیل توانایی در شناسایی روابط بین متغیرهای پیش‌بینی کننده و یک خروجی دودویی، انتخاب شد (مطیعی و قاسم نژاد، ۱۳۹۶).

مراحل تحلیل داده، شامل پیش‌پردازش، توسعه مدل، اعمال وزن‌ها و ارزیابی عملکرد، با بهره‌گیری از نرم‌افزار آماری JMP Pro نسخه ۱۷،۲،۰ انجام گرفت. این نرم‌افزار که توسط شرکت SAS Institute Inc. توسعه یافته، به عنوان یکی از ابزارهای قدرتمند در حوزه تحلیل و اکتشاف داده‌های آماری شناخته می‌شود. دلیل انتخاب این پلتفرم، توانمندی‌های قابل توجه آن در مدلسازی رگرسیونی، مدیریت داده‌های نامتوازن از طریق اختصاص وزن، و همچنین ارائه ابزارهای تحلیلی بصری و جامع برای ارزیابی مدل‌ها بوده است. افزون بر این، ترکیب محیط کاربری گرافیکی تعاملی با روش‌های پیشرفته آماری و الگوریتم‌های یادگیری ماشین، این نرم‌افزار را به گزینه‌ای مناسب و پرکاربرد برای تحلیل داده‌های پیچیده در میان پژوهشگران و متخصصان تبدیل کرده است.

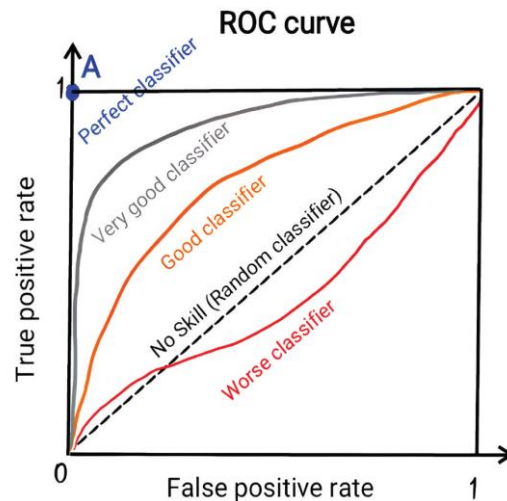
(Gross and Harris, 2013). برای اطمینان از قابلیت تعمیم‌پذیری مدل و جلوگیری از بیش‌برازش، کل مجموعه داده به صورت تصادفی به دو زیرمجموعه تقسیم شد؛ ۷۰٪ برای آموزش^{۱۹} مدل و ۳۰٪ برای اعتبارسنجی^{۲۰}. مدل به طور کامل روی مجموعه آموزشی، آموزش داده شد و عملکرد آن بر روی مجموعه اعتبارسنجی که مدل آنها را ندیده بود، صحت‌سنجی گردید.

۴-۲- معیارهای ارزیابی مدل

منحنی مشخصه عملکرد سیستم^{۲۱} (ROC) به عنوان یک ابزار کمی در ارزیابی مسائل طبقه‌بندی به کار می‌رود و امکان سنجش دقیق توان تشخیص یک سامانه را فراهم می‌سازد. در این رویکرد، عملکرد مدل بر پایه شاخص‌هایی ارزیابی می‌شود که وابسته به قضاوت‌های فردی نبوده و بر معیارهای عینی استوار هستند؛ از این رو، تحلیل ارائه شده از دقت و قابلیت تفکیک‌پذیری مدل، قابل اعتمادتر خواهد بود (Chang and Newman., 2024)

کارایی الگوریتم‌های دسته‌بندی دو دویی معمولاً به وسیله شاخص‌هایی مثل حساسیت^{۲۲} و ویژگی^{۲۳} سنجیده می‌شوند. اما در نمودار ROC هر دوی این شاخص‌ها ترکیب شده و به صورت یک منحنی نمایش داده می‌شوند. اغلب برای بررسی کارایی الگوریتم‌های دسته‌بندی یا ایجاد داده‌های رسته‌ای از منحنی ROC استفاده می‌کنند. این موضوع در شاخه یادگیری ماشین با نظارت، بیشتر مورد توجه قرار گرفته است.

منحنی ROC با رسم نرخ مثبت واقعی^{۲۴} (TPR) در مقابل نرخ مثبت کاذب^{۲۵} (FPR) در سطوح مختلف آستانه تصمیم‌گیری شکل می‌گیرد. در این چارچوب، TPR بیانگر نسبت پیش‌بینی‌های صحیح از وقوع شکست به کل رخداد‌های واقعی است و با عناوینی نظیر حساسیت، بازیابی یا احتمال کشف نیز شناخته می‌شود. در مقابل، FPR نشان‌دهنده سهم پیش‌بینی‌های نادرست از وقوع شکست در میان مواردی است که در واقع شکست رخ نداده است و معمولاً به عنوان احتمال هشدار کاذب در نظر گرفته می‌شود؛ این شاخص را می‌توان به صورت (۱ منهای ویژگی) نیز بیان کرد. در شکل ۳، منحنی ROC ارتباط میان نرخ مثبت واقعی به عنوان منفعت و نرخ مثبت کاذب به عنوان هزینه را به خوبی نمایش می‌دهد (Salehinasab and Goshtasbi Rad, 2025).



شکل ۳- نمودار منحنی ROC (Salehinasab and Goshtasbi Rad, 2025)

با توجه به اینکه نرخ مثبت واقعی (TPR) معادل حساسیت و نرخ مثبت کاذب (FPR) برابر با (ویژگی-۱) است، نمودار ROC در برخی منابع به صورت نمودار حساسیت در برابر (ویژگی-۱) نیز معرفی می شود. هر خروجی پیش بینی یا هر حالت از ماتریس درهم ریختگی، به صورت یک نقطه در فضای ROC قابل نمایش است.

در حالت ایده آل، بهترین مدل پیش بینی نقطه ای در گوشه بالای سمت چپ این فضا، یعنی مختصات (۰، ۱)، تولید می کند که بیانگر حساسیت کامل (عدم وجود منفی کاذب) و ویژگی کامل (عدم وجود مثبت کاذب) است؛ این وضعیت به عنوان طبقه بندی بی نقص شناخته می شود. در مقابل، عملکرد تصادفی به صورت نقاطی روی خط قطری از گوشه پایین چپ تا گوشه بالا راست ظاهر می شود که فضای ROC را به دو بخش تقسیم می کند. نقاطی که بالاتر از این خط قرار دارند نشان دهنده عملکردی بهتر از حد تصادفی هستند، در حالی که نقاط زیر آن بیانگر عملکرد ضعیف تر از تصادف اند. جالب آن که اگر یک مدل به طور مداوم نتایج نادرست ارائه دهد، با معکوس سازی خروجی های آن می توان به عملکردی قابل قبول دست یافت. از منحنی ROC می توان برای ارزیابی کارایی مدل های رگرسیون لجستیک نیز بهره گرفت. در این زمینه، معیار مساحت زیر منحنی (AUC) شاخصی مهم تلقی می شود؛ به گونه ای که هرچه مقدار آن بزرگ تر باشد، نشان دهنده توان تفکیک پذیری و عملکرد بهتر مدل خواهد بود (Li et al. 2019).

در مسئله پیش بینی دو کلاسه ای (کلاسه بندی دودویی) خروجی طبقه بندی، شامل ۴ نوع نتیجه است. به صورتی که دسته اول، همان پیش بینی های صحیح هستند که شامل مثبت صحیح (TP) و منفی صحیح (TN) می شوند و دسته دوم نیز همان پیش بینی های نادرست هستند که شامل مثبت نادرست (FP) و منفی نادرست (FN) می شوند و می توان آن ها را در قالب ماتریسی که در جدول ۴ نشان داده شده است (که به صورت 2×2 می باشد) ارائه کرد (Wu and Liu, 2017).

جدول ۴- نتایج پیش‌بینی دو کلاسه‌ای (ماتریس برهم ریختگی^{۲۷})

Status	Predicted Positive	Predicted Negative
Actual Positive	True Positives (TP)	False Negatives (FN)
Actual Negative	False Positives (FP)	True Negatives (TN)

برای ارزیابی عملکرد مدل سازی ناشی از حوادث به وقوع پیوسته در شبکه توزیع آب مورد مطالعه، از روابط ۱ تا ۱۵ استفاده شده است. هرچه مقدار این معادلات بیشتر باشد، به معنای عملکرد بهتر مدل و بالاتر بودن قدرت مدل در پیش‌گویی و دسته‌بندی داده‌ها می‌باشد. دقت کلی^{۲۸}، حساسیت، ویژگی، دقت^{۲۹} (صحت) و F-Score از جمله معادلاتی هستند که برای انجام ارزیابی مدل سازی‌ها، از محبوبیت و کارایی بالاتری در میان پژوهشگران برخوردارند. در این میان برخی از مفاهیم دارای معادلات یکسانی می‌باشند که تمامی آن‌ها در کنار هم قرار داده شده‌اند و معادلات آن‌ها نیز به شرح زیر می‌باشد (Wu and Liu, 2017).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (۱)$$

$$\text{Sensitivity} = \text{Recall} = \text{True Positive Rate} = \frac{TP}{TP+FN} \quad (۲)$$

$$\text{Specificity} = \text{True Negative Rate} = \frac{TN}{TN+FP} \quad (۳)$$

$$\text{Precision} = \text{Positive Predictive Value} = \frac{TP}{TP+FP} \quad (۴)$$

$$\text{F-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۵)$$

۳- نتایج و بحث

در این بخش، یافته‌های حاصل از فرآیند توسعه و ارزیابی مدل رگرسیون لجستیک وزن‌دار برای پیش‌بینی شکست لوله ارائه شده است. نتایج در شش زیربخش اصلی ارائه گردید:

۳-۱- فرآیند توسعه و انتخاب مدل نهایی

در مرحله اولیه مدل‌سازی، تمام متغیرهای بالقوه مهندسی و فیزیکی شامل طول، سن، جنس، قطر، فشار و عمق نصب لوله در مدل رگرسیون لجستیک گنجانده شد. این رویکرد مبتنی بر این فرضیه بود که عمق نصب می‌تواند از طریق تأثیرگذاری بر بار خاک و تنش‌های محیطی، بر پایداری لوله مؤثر باشد. با این حال، نتایج اولیه مدل نشان داد که متغیر عمق نصب هیچ‌گونه تأثیر آماری معناداری بر احتمال شکست ندارد ($P\text{-Value} = 0.79$, $\text{LogWorth} = 0.101$). لازم به ذکر است برای معنادار بودن یک متغیر مقدار $P\text{-Value}$ باید کمتر از ۰/۰۵ باشد و هرچه به صفر نزدیکتر که در نتیجه LogWorth عددی بزرگتر خواهد بود، نشان دهنده

معناداری بیشتر متغیر است. به منظور رعایت اصل سادگی مدل و توسعه یک مدل بهینه و قابل تفسیر، این متغیر غیرمعنادار از مدل نهایی حذف شد. اصل سادگی، بیان می‌کند که در میان مدل‌هایی با عملکرد مشابه، مدل ساده‌تر (با متغیرهای کمتر) ترجیح داده می‌شود. حذف متغیرهای بی‌اثر نه تنها مدل را ساده‌تر می‌کند، بلکه با کاهش پارامترهای غیر ضروری، خطر بیش‌برازش را نیز کاهش می‌دهد (Burnham and Anderson, 2002). در نتیجه، مدل نهایی شامل پنج متغیر پیش‌بینی‌کننده معنادار بود: طول، سن، جنس، قطر و فشار لوله.

۳-۲- ارزیابی کلی و معناداری مدل نهایی

معیارهای برازش برای مدل نهایی ۵ متغیره مانند آزمون نسبت در ست‌نمایی^{۳۰} نشان می‌دهد که مدل به طور کلی بسیار معنادار است ($\text{ChiSq} < 0.0001$) و عملکرد آن به مراتب بهتر از یک مدل صرف (مدل بدون هیچ متغیر پیش‌بینی‌کننده) است. همچنین معیار RSquare (U) برابر با ۰/۳۵ است که نشان می‌دهد مدل حدود ۳۵٪ از تغییرات در لگاریتم احتمالات شکست را توضیح می‌دهد. این مقدار برای مسائل مهندسی واقعی که با داده‌های پرنویز همراه هستند، یک مقدار قابل قبول محسوب می‌شود. مقادیر بین ۰/۲ تا ۰/۴ برای این معیار، عموماً به عنوان یک برازش خوب در نظر گرفته می‌شوند (Hosmer et al, 2013).

۳-۳- تحلیل اهمیت و تأثیر متغیرهای پیش‌بینی‌کننده

جدول ۵، اهمیت آماری هر متغیر را بر اساس LogWorth و P-Value نشان می‌دهد. نتایج رتبه‌بندی زیر را برای متغیرها مشخص می‌کند: قطر لوله > فشار > جنس لوله > سن لوله > طول لوله

جدول ۵: اهمیت آماری متغیرهای پیش‌بینی‌کننده در مدل نهایی

متغیر	LogWorth	P-Value
Len(m)	407.191	<.0001
age(year)	99.784	<.0001
Material	29.615	<.0001
Pr (m H ₂ O)	5.77	<.0001
Dia (mm)	4.778	<.0001

جدول ۶ ضرایب تخمین زده شده برای هر متغیر را نشان می‌دهد. لازم به ذکر است نرم افزار به صورت خودکار یکی از متغیرها (نه بهترین و نه بدترین) را به عنوان مرجع در نظر می‌گیرد. در این پژوهش برای متغیر دسته‌بندی شده جنس، UPVC به عنوان گروه مرجع در نظر گرفته شده است. نسبت شانس برای جنس یعنی متغیر دیگر به چه نسبتی از متغیر مرجع (با ثابت بودن سایر متغیرها) دارای شانس شکست است. به عنوان نمونه لوله آریست ۳/۳۸ برابر بیشتر از لوله UPVC با شکست مواجه خواهد شد. نسبت شانس برای سایر متغیرها نیز میزان اهمیت آن متغیر برای شکست است که هرچه عدد بزرگتر از ۱ باشد یعنی تأثیر مثبت

در شکست و در صورتیکه عدد کمتر از ۱ باشد یعنی تاثیر منفی در شکست، خواهد داشت.

جدول ۶: ضرایب تخمین زده شده مدل نهایی

متغیر	ضریب (Estimate)	نسبت شانس (Odds Ratio)	P-Value
Intercept	1.429	-	<.0001
Material[AC]	1.218	3.38	<.0001
Material[DI]	-1.306	0.27	<.0001
Material[PE]	-0.858	0.42	<.0001
Material[ST]	-0.874	0.42	0.122
Len(m)	0.022	1.02	<.0001
Pr (m H ₂ O)	0.011	1.01	<.0001
Dia (mm)	-0.0024	0.998	<.0001
age(year)	-0.104	0.9	<.0001

جنس لوله: نتایج نشان می‌دهد که با ثابت نگه داشتن سایر شرایط، لوله‌های آزیست سیمانی (AC) به طور قابل توجهی مستعدتر به شکست نسبت به لوله‌های UPVC هستند (شانس شکست ۳/۴ برابر). در مقابل، لوله‌های چدن داکتیلی (DI) و پلی‌اتیلن (PE) به عنوان بادوام‌ترین گزینه‌ها شناسایی شدند که شانس شکست آن‌ها به ترتیب ۷۳٪ و ۵۸٪ کمتر از لوله‌های UPVC است.

طول لوله: ضریب مثبت و معنادار (۰/۰۲۲) نشان می‌دهد که با افزایش طول لوله، احتمال شکست نیز افزایش می‌یابد.

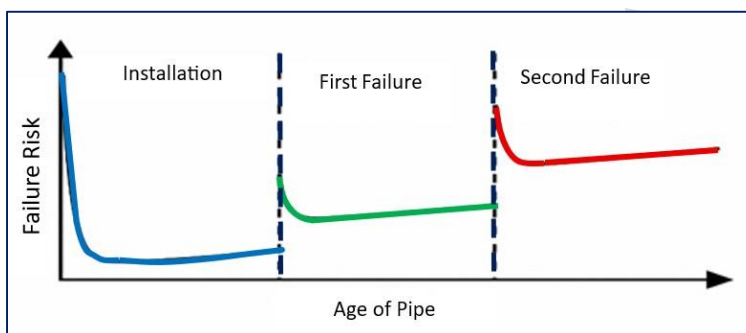
فشار: ضریب مثبت (۰/۰۱۱) نشان‌دهنده تأثیر مثبت فشار بر احتمال شکست است.

قطر لوله: ضریب منفی (۰/۰۲۴-) نشان می‌دهد که لوله‌های با قطر بزرگتر، احتمال شکست کمتری دارند.

سن لوله: جالب‌ترین یافته، ضریب منفی و بسیار معنادار (۰/۱۰۴-) برای سن لوله است. این نتیجه نشان می‌دهد که با افزایش سن لوله، احتمال شکست کاهش می‌یابد.

یکی از قابل توجه‌ترین یافته‌های این پژوهش، رابطه منفی و معنادار سن لوله با احتمال شکست است. این نتیجه در تضاد با فرضیه کلاسیک فرسودگی به نظر می‌رسد، اما می‌توان آن را به خوبی با پدیده سوگیری بازماندگان^{۳۱} توجیه کرد. داده‌های این مطالعه نشان می‌دهد که بخش قابل توجهی از شکست‌ها در سنین پایین و یا تکرار شکست‌ها در مدتی کوتاه (به عنوان مثال برای یکی از لوله‌ها در مدت یک ماه، ۳ شکست ثبت شد) رخ می‌دهد که این امر می‌تواند ناشی از مشکلات کیفیت ساخت یا نصب نامناسب باشد. در نتیجه، لوله‌هایی که به سنین بالاتر می‌رسند، نماینده گروهی از بازماندگان مقاوم هستند که ذاتاً مستعد شکست کمتری هستند. مدل رگرسیون، این ویژگی انتخابی را به اشتباه به عنوان یک اثر محافظتی از افزایش سن تفسیر می‌کند. این یافته با نتایج

مطالعات دیگر در زمینه قابلیت اطمینان زیرساخت‌ها مطابقت دارد که نشان می‌دهند نرخ شکست لوله‌ها اغلب یک منحنی U شکل است: در ابتدا به دلیل مشکلات اولیه بالا، سپس کاهش یافته و در نهایت به دلیل فرسودگی مجدداً افزایش می‌یابد (Kleiner and Rajani, 2001). همچنین رابطه منفی سن لوله و احتمال شکست را می‌توان با مدل نظری عمر شکست لوله که توسط Scheidegger et al. (2015) ارائه شده است، به خوبی توجیه کرد. شکل ۴، منحنی نظری ریسک شکست در طول عمر یک لوله را به تصویر می‌کشد. این مدل سه فاز اصلی را پیش‌بینی می‌کند: فاز اول با ریسک بالا به دلیل نقص‌های اولیه، فاز دوم با ریسک پایین و پایدار (عمر مفید)، و فاز سوم با ریسک فزاینده به دلیل فرسودگی.



شکل ۴- ریسک نظری شکست در طی عمر لوله (Scheidegger et al. 2015)

ضریب منفی سن در این پژوهش، مستقیماً بازتاب‌دهنده پدیده سوگیری بازماندگان است که در فاز اول رخ می‌دهد. شکست‌های زودهنگام لوله‌های ضعیف، منجر به باقی ماندن یک جمعیت بازمانده مقاوم‌تر در سنین بالاتر می‌شود. علاوه بر این، شکل ۴، اهمیت دو چالش دیگری را که در این پژوهش به آن‌ها پرداخته شد، برجسته می‌سازد: (۱) لزوم در نظر گرفتن شکست‌های زودهنگام که منجر به سوگیری بازماندگان می‌شود و (۲) لزوم تحلیل شکست‌های تکرار شونده، زیرا همانطور که در شکل مشخص است، هر تعمیر، ریسک پایه لوله را به طور دائمی افزایش می‌دهد.

همانطور که در جدول ۳ آورده شد حداکثر سن لوله‌ها در شبکه مورد بررسی در این تحقیق، ۴۵ سال و میانگین آن‌ها ۲۳/۸ سال است لذا می‌توان نتیجه گرفت اکثر لوله‌های موجود در شبکه هنوز وارد فاز فرسودگی نشده‌اند. بنابراین، ضریب منفی سن در این مدل، نه یک خطا، بلکه یک شاخص قدرتمندی از اهمیت حیاتی کیفیت نصب و اجرای اولیه لوله و کنترل کیفیت آن در سال‌های اولیه عمر لوله است.

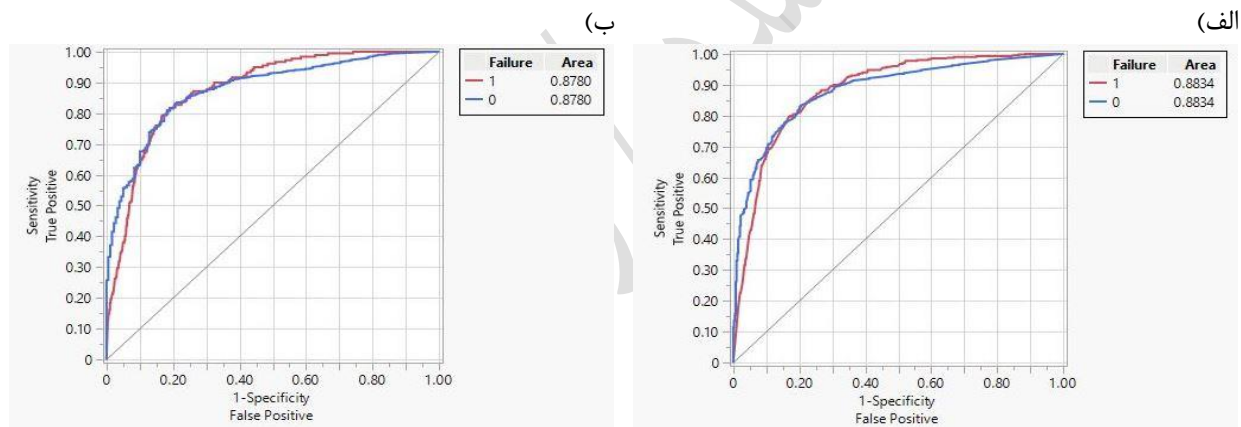
۳-۴- ارزیابی عملکرد پیش‌بینی‌کننده مدل

برای ارزیابی قابلیت تعمیم‌پذیری مدل، عملکرد آن بر روی مجموعه‌های آموزش و اعتبارسنجی سنجیده شد. جدول ۷ نتایج این ارزیابی را بر اساس معیارهای استخراج‌شده از ماتریس درهم‌ریختگی نشان می‌دهد.

جدول ۷: ارزیابی قابلیت تعمیم‌پذیری مدل

معیار	مجموعه آموزشی	مجموعه اعتبارسنجی
حساسیت (Sensitivity)	۰/۷۸	۰/۷۶۵
ویژگی (Specificity)	۰/۸۴	۰/۸۳۴
دقت (Precision)	۰/۴۳۶	۰/۴۴۲
دقت کلی (Accuracy)	۰/۸۳۲	۰/۸۲۸
امتیاز-F1 (F1-Score)	۰/۵۶۹	۰/۵۶

نتایج نشان‌دهنده عملکرد بسیار پایدار مدل است. تفاوت جزئی معیارها بین مجموعه آموزش و اعتبارسنجی (مثلاً کاهش حساسیت از ۰/۷۸ به ۰/۷۶۵) حاکی از عدم وجود بیش‌برازش و توانایی بالای مدل در تعمیم نتایج به داده‌های جدید است. مهم‌ترین معیار برای این کاربرد، حساسیت است که نشان می‌دهد مدل قادر به شناسایی حدود ۷۶٪/۵ از لوله‌های واقعاً در معرض خطر در مجموعه اعتبارسنجی بوده است. این معیار برای کاربردهای پیشگیرانه بسیار حیاتی است. منحنی ROC (شکل ۵) مساحت زیر منحنی (AUC) برابر با ۰/۸۸ را برای مجموعه اعتبارسنجی نشان می‌دهد که قدرت تفکیک‌کنندگی عالی مدل را تأیید می‌کند.



شکل ۵- منحنی ROC-الف) براساس داده‌های آموزشی؛ ب) براساس داده‌های صحت‌سنجی

۵-۳-تفسیر یافته‌های کلیدی

نتایج این پژوهش، یک مدل رگرسیون لجستیک وزن‌دار با عملکرد قابل قبول برای پیش‌بینی شکست لوله ارائه داد. مدل با به دست آوردن AUC برابر با ۰/۸۸ در مجموعه اعتبارسنجی، قدرت تفکیک‌کنندگی عالی از خود نشان داد و توانایی خود را در شناسایی لوله‌های در معرض خطر به اثبات رساند. تحلیل اهمیت متغیرها، طول و سن لوله را به ترتیب به عنوان مهم‌ترین عوامل پیش‌بینی‌کننده شناسایی کرد.

جالب‌ترین و قابل توجه‌ترین یافته این مطالعه، رابطه منفی و معنادار بین سن لوله و احتمال شکست بود (ضریب = -۰/۱۰۴). این

نتیجه در نگاه اول با فرضیه کلاسیک فرسودگی مواد در تضاد است. با این حال، این پدیده را می‌توان به خوبی با مفهوم آماری سوگیری بازماندگان توجیه کرد. این سوگیری زمانی رخ می‌دهد که لوله‌های ضعیف‌تر، که به دلیل کمبودهای کیفی (مانند نصب نامناسب یا نقص در ساخت) مستعد شکست هستند، در سنین پایین از شبکه خارج می‌شوند. در نتیجه، جمعیت لوله‌های باقی‌مانده در سنین بالاتر، دیگر یک نمونه تصادفی نیستند، بلکه گروهی از بازماندگان مقاوم هستند که ذاتاً مستعد شکست کمتری هستند.

این یافته با مدل نظری عمر شکست لوله که توسط Scheidegger et al. (2015) ارائه شده است، کاملاً سازگار است. این مدل سه فاز را پیش‌بینی می‌کند: فاز اول با ریسک بالا به دلیل نقص‌های اولیه، فاز دوم با ریسک پایین (عمر مفید) و فاز سوم با ریسک فزاینده به دلیل فرسودگی. مدل پژوهش حاضر به احتمال زیاد، در فاز دوم یا اوایل فاز سوم در حال تحلیل داده‌هاست که در آن تأثیر بازماندگان مقاوم بر تأثیر تدریجی فرسودگی غلبه کرده است. این یافته، پیامی روشن و بسیار مهم برای مدیران شبکه دارد: کیفیت نصب و کنترل کیفیت در سال‌های اولیه عمر لوله، به اندازه کیفیت خود لوله حیاتی است.

در مورد سایر متغیرها، نتایج کاملاً با اصول مهندسی و یافته‌های مطالعات پیشین همخوانی دارد. تأثیر مثبت طول و فشار و تأثیر منفی قطر بر احتمال شکست، کاملاً قابل توجیه است. لوله‌های بلندتر در معرض قرار گرفتن بیشتری دارند و فشار بالاتر تنش‌های داخلی را افزایش می‌دهد. در مقابل، لوله‌های با قطر بزرگتر معمولاً دیواره‌های ضخیم‌تری دارند که آن‌ها را در برابر شکست مقاوم‌تر می‌کند (Yamijala et al. 2009).

۳-۶- پیامدهای عملی برای مدیریت شبکه توزیع آب

نتایج این پژوهش کاربردهای عملی مستقیمی برای مدیران و بهره‌برداران شبکه‌های توزیع آب دارد که می‌توان به صورت زیر آن‌ها را بر شمرد:

الف) اولویت‌بندی مبتنی بر ریسک: مدل توسعه‌یافته، ابزاری قدرتمند برای ایجاد نقشه ریسک شکست لوله در سراسر شبکه فراهم می‌کند. مدیران می‌توانند با استفاده از خروجی احتمال مدل، لوله‌ها را بر اساس سطح ریسک اولویت‌بندی کرده و منابع محدود بازرسی، تعمیر و جایگزینی را به مؤثرترین شکل تخصیص دهند.

ب) تمرکز بر کیفیت نصب: با توجه به یافته کلیدی در مورد تأثیر منفی سن و اهمیت شکست‌های زود هنگام، می‌توان گفت قابل نتیجه‌گیری است که کیفیت نصب و بازرسی اولیه به اندازه کیفیت خود لوله اهمیت دارد. سرمایه‌گذاری در فرآیندهای نصب استاندارد و کنترل کیفیت، می‌تواند به طور قابل توجهی نرخ شکست را در بلندمدت کاهش دهد.

ج) مدیریت متمایز بر اساس جنس لوله: نتایج نشان می‌دهد که استراتژی‌های مدیریتی باید بر اساس جنس لوله متفاوت باشند. برای لوله‌های AC و UPVC، برنامه‌های بازرسی فشرده‌تر و زود هنگام ضروری است، در حالی که برای لوله‌های DI و PE می‌توان از بازرسی‌های با فاصله زمانی طولانی‌تر استفاده کرد.

۴- نتیجه‌گیری

در این بخش، نتایج کلی، محدودیت‌ها و پیشنهادات برای تحقیقات آتی ارائه شده است.

۴-۱- بحث و نتیجه‌گیری

این پژوهش با درک عمیق از چالش‌های ذاتی داده‌های شبکه‌های توزیع آب و با هدف غلبه بر آنها در پیش‌بینی شکست لوله‌های آب شهری، از جمله عدم تعادل کلاس‌ها و رویدادهای تکرارشونده، به توسعه و ارزیابی یک مدل یادگیری ماشین جامع و تفسیرپذیر پرداخت. با بهره‌گیری از مجموعه داده‌های واقعی شبکه توزیع آب تهران، مدلی مبتنی بر رگرسیون لجستیک وزن‌دار با موفقیت توسعه یافت که توانایی بالایی در شناسایی لوله‌های در معرض خطر از خود نشان داد.

نتایج کلیدی این مطالعه نشان‌دهنده قدرت تفکیک‌کنندگی عالی مدل، با AUC برابر با ۰/۸۸ در داده‌های اعتبار سنجی و همچنین عملکرد پایدار آن در مواجهه با داده‌های جدید است. تحلیل اهمیت متغیرها، طول لوله و سن لوله را به عنوان مهم‌ترین عوامل پیش‌بینی‌کننده معرفی کرد. یافته جالب توجه در این تحقیق، کشف رابطه منفی بین سن لوله و احتمال شکست بود که با مفهوم سوگیری بازماندگان و کیفیت نصب و نگهداری لوله‌های قدیمی‌تر توجیه شد. این امر، اهمیت فراتر رفتن از صرف ویژگی‌های فیزیکی و در نظر گرفتن سایر عوامل مؤثر را برجسته می‌سازد.

مهم‌ترین نوآوری این تحقیق، ارائه یک چارچوب مدلسازی جامع است که به طور همزمان سه رویکرد کلیدی را در بر می‌گیرد: مدیریت هوشمندانه داده‌های نامتعادل، مدل‌سازی شکست‌های تکرارشونده، و اعتبار سنجی قوی. این یکپارچگی، مدلی را به ارائه داده که نه تنها از نظر علمی مستحکم است، بلکه قابلیت پیاده‌سازی عملی در مقیاس وسیع را نیز دارد.

این مدل، ابزاری علمی و کاربردی برای مدیران شبکه‌های آبرسانی فراهم می‌آورد تا بتوانند از رویکردهای سنتی واکنشی به سمت نگهداری پیشگیرانه و مبتنی بر ریسک حرکت کنند. این گذار، نه تنها به کاهش هزینه‌های عملیاتی و افزایش بهره‌وری منجر می‌شود، بلکه با ارتقاء پایداری و اطمینان بخشی به سرویس‌دهی، منافع اجتماعی و اقتصادی گسترده‌ای را به همراه خواهد داشت.

۲-۴- محدودیت‌های مطالعه و پیشنهادات برای تحقیقات آتی

این پژوهش با تکیه بر داده‌های گردآوری شده از یک ناحیه مشخص در شهر تهران انجام شده است؛ بنابراین تعمیم نتایج به سایر مناطق با ویژگی‌های متفاوت اقلیمی، خصوصیات خاک و الگوهای مصرف، مستلزم بررسی و اعتبارسنجی‌های تکمیلی است. همچنین برخی عوامل اثرگذار نظیر جنس و ویژگی‌های خاک، میزان بار ترافیکی و کیفیت آب در مدل لحاظ نشده‌اند. علاوه بر این باید این نکته مهم را در نظر داشت که با توجه به اینکه فشار شبکه در طول زمان (سال‌های مختلف) ثابت نبوده و به دلایل مختلف دارای تغییراتی خواهد بود، برای تعیین فشار به طور دقیق در زمان شکست، نیازمند پایش و کنترل در لحظه^{۳۲} شبکه خواهد بود که این پارامتر به عنوان یک داده در اختیار محققین قرار داده نشده بود. در نظر گرفتن این متغیرها در مطالعات آتی می‌تواند به بهبود دقت و کارایی مدل منجر شود.

بعد زمانی مطالعه: مدل پیشنهادی بر پایه داده‌های مربوط به یک دوره ۱۲ ساله توسعه یافته است. با این حال، برای دستیابی به درک جامع‌تری از چرخه عمر لوله‌ها، استفاده از داده‌های بلندمدت‌تر ضروری به نظر می‌رسد؛ چراکه این امر امکان مدل‌سازی دقیق‌تر مرحله نهایی عمر، یعنی فاز فرسودگی، را فراهم می‌کند.

سپاسگزاری

نویسندگان تشکر و سپاس خود را از سردبیر و داوران محترم که با ارائه نظرات خود باعث بهبود و ارتقای علمی این مقاله شدند و همچنین شرکت آب و فاضلاب منطقه یک تهران جهت همکاری در تهیه داده‌های مورد نیاز تحقیق ابراز می‌دارند.

نویسندگان هیچگونه تعارض منافع با هم ندارند.

-پی‌نوشت‌ها

- 1- Machine Learning
- 2- Regression Analysis
- 3- Decision Tree
- 4- Random Forest
- 5- Support Vector Machine
- 6- Extreme Gradient Boosting
- 7- Logistic Regression
- 8- Support Vector Classification
- 9- Overfitting
- 10- Binary Classifier
- 11- Ductile Iron
- 12- Polyethylene

- 13- Asbestos- Cement
- 14- Steel
- 15- Un Plasticized Polyvinyl Chloride
- 16- Failure
- 17- Event-Time
- 18- Re-weighting
- 19- Training
- 20- Validation
- 21- Receiver Operating Characteristic
- 22- Sensitivity
- 23- Specificity
- 24- True Positives Rate
- 25- False Positives Rate
- 26- Confusion Matrix
- 27- Area Under the Curve
- 28- Accuracy
- 29- Precision
- 30- Likelihood Ratio Test
- 31- Survivorship Bias
- 32- Real-Time

۵-مراجع

نوذری، ن، مستوری، ر، میرحسینی، س.م. (۱۴۰۲). بررسی کج شدن ناحیه‌ای درختان و سازه‌ها ناشی از نشست خاک در اثر شکستگی لوله شبکه توزیع آب در تهران. علوم و مهندسی آب و فاضلاب، ۸(۲)، ۳۷-۴۶.

<https://doi.org/10.22112/jwwse.2022.348911.1319>

مصلحی، ا، جلیلی قاضی زاده، م.ر. (۱۳۹۵). مروری بر روابط فشار- شکستگی در شبکه‌های آبرسانی. علوم و مهندسی آب و فاضلاب، ۱۱(۱)، ۱۹-۱۱.

<https://doi.org/10.22112/jwwse.2017.51053>

مطیعی، ه و قاسم نژاد، س. (۱۳۹۶). کاربرد و توسعه مدل‌های رگرسیونی برای پیش‌بینی میزان شکست لوله‌های شبکه توزیع آب شهری-مورد مطالعاتی ناحیه یک منطقه یک تهران. علوم و مهندسی آب و فاضلاب، ۲(۲)، ۴۸-۵۸.

<https://doi.org/10.22112/jwwse.2017.89374.1024>

Asadi, Y. (2024). Employing machine learning in water infrastructure management: predicting pipeline failures for improved maintenance and sustainable operations. Industrial artificial intelligence. 2(1). 8. <https://doi.org/10.1007/s44244-024-00022-w>

Barton, N.A., Hallett, S.H., Jude, S.R. and Tran, T.H. (2022). An evolution of statistical pipe failure models for drinking water networks: A targeted review. Water Supply, 22(4), 3784-3813. <https://doi.org/10.2166/ws.2022.019>

Burnham, K. P. and Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-

- theoretic approach (2nd ed.). Springer-Verlag. <https://doi.org/10.1007/b97636>
- Chang, P.W. and Newman, T.B. (2024). Receiver operating characteristic (ROC) curves: the basics and beyond. *Hospital pediatrics*, 14(7), pp.e330-e334. <https://doi.org/10.1542/hpeds.2023-007462>
- Chen, R., Wang, Q. and Javanmardi, A. (2025). A Review of the Application of Machine Learning for Pipeline Integrity Predictive Analysis in Water Distribution Networks. *Archives of Computational Methods in Engineering*. 32(6). 3821-3849. <https://doi.org/10.1007/s11831-025-10251-6>
- Farmani, R., Kakoudakis, K., Behzadian, K. and Butler, D. (2017). Pipe failure prediction in water distribution systems considering static and dynamic factors. *Procedia Engineering*. 186. 117-126. <https://doi.org/10.1016/j.proeng.2017.03.217>
- Gross, R.E. and Harris, S.P. (2013). Statistical performance evaluation of soft (elastomer) seat pressure relief valves. In *Pressure Vessels and Piping Conference* (Vol. 55706, p. V06AT06A077). American Society of Mechanical Engineers. <https://doi.org/10.1115/PVP2013-97031>
- He, H. and Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*. 21(9). 1263-1284. <https://doi.org/10.1109/TKDE.2008.239>
- Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X. (2013). *Applied logistic regression*. John Wiley & Sons. <https://doi.org/10.1002/0471722146>
- Kakoudakis, K., Behzadian, K., Farmani, R. and Butler, D. (2017). Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering. *Urban Water Journal*. 14(7). 737-742. <https://doi.org/10.1080/1573062X.2016.1253755>
- Karadirek, I.E., Kaya-Basar, E. and Akdeniz, T. (2024). A study on pipe failure analysis in water distribution systems using logistic regression. *Water Supply*. 24(1). 176-186. <https://doi.org/10.2166/ws.2023.335>
- Karimian, F., Kaddoura, K., Zayed, T., Hawari, A. and Moselhi, O. (2021). Prediction of breaks in municipal drinking water linear assets. *Journal of Pipeline Systems Engineering and Practice*. 12(1) 04020060. [https://doi.org/10.1061/\(ASCE\)PS.1949-1204.0000511](https://doi.org/10.1061/(ASCE)PS.1949-1204.0000511)
- Kleiner, Y. and Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban water*. 3(3).131-150. [https://doi.org/10.1016/S1462-0758\(01\)00033-4](https://doi.org/10.1016/S1462-0758(01)00033-4)
- Latifi, M., Beig Zali, R., Javadi, A.A. and Farmani, R. (2024). Customised-sampling approach for pipe failure prediction in water distribution networks. *Scientific Reports*. 14(1).18224. <https://doi.org/10.1038/s41598-024-69109-9>
- Leštáková, M., Logan, K. T., Rehm, I. S., Pelz, P. F. and Friesen, J. (2024). Do resilience metrics of water distribution systems really assess resilience? A critical review. *Water research*. 248. 120820. <https://doi.org/10.1016/j.watres.2023.120820>
- Li, C., Chen, J., & Qin, G. (2019). Partial Youden index and its inferences. *Journal of biopharmaceutical statistics*, 29(2), 385-399. <https://doi.org/10.1080/10543406.2018.1535502>
- Omar, A., Delnaz, A. and Nik-Bakht, M. (2023). Comparative analysis of machine learning techniques for predicting water main failures in the City of Kitchener. *Journal of Infrastructure Intelligence and Resilience*. 2(3). 100044. <https://doi.org/10.1016/j.iintel.2023.100044>
- Rifaai, T.M., Abokifa, A.A. and Sela, L. (2022). Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems. *Reliability engineering & system safety*, 220, 108271. <https://doi.org/10.1016/j.ress.2021.108271>
- Robles-Velasco, A., Cortés, P., Muñuzuri, J. and Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliability Engineering & System Safety*.

196. 106754. <https://doi.org/10.1016/j.res.2019.106754>
- Salehinasab, S. and Goshtasbi Rad, E. (2025). Enhancing pipe failure prediction using cluster-specific classification via K-means clustering. Sustainable and Resilient Infrastructure.1-30. <https://doi.org/10.1080/23789689.2025.2470504>
- Scheidegger, A., Leitao, J.P. and Scholten, L. (2015). Statistical failure models for water distribution pipes—A review from a unified perspective. Water research. 83. 237-247. <https://doi.org/10.1016/j.watres.2015.06.027>
- Shirzad, A. and Safari, M.J.S. (2019). Pipe failure rate prediction in water distribution networks using multivariate adaptive regression splines and random forest techniques. Urban Water Journal, 16(9),653-661. <https://doi.org/10.1080/1573062X.2020.1713384>
- Snider, B. and McBean, E.A. (2020). Improving urban water security through pipe-break prediction models: Machine learning or survival analysis. Journal of Environmental Engineering. 146(3). 04019129. [https://doi.org/10.1061/\(ASCE\)EE.1943-7870.0001657](https://doi.org/10.1061/(ASCE)EE.1943-7870.0001657)
- Solomatine, D.P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. Journal of hydroinformatics, 10(1), 3-22. <https://doi.org/10.2166/hydro.2008.015>
- Wang, X., Mazumder, R.K., Salarieh, B., Salman, A.M., Shafieezadeh, A. and Li, Y. (2022). Machine learning for risk and resilience assessment in structural engineering, Progress and future trends. Journal of Structural Engineering. 148(8). 03122003. [https://doi.org/10.1061/\(ASCE\)ST.1943-541X.0003392](https://doi.org/10.1061/(ASCE)ST.1943-541X.0003392)
- Wu, Y. and Liu, S. (2017). "A review of data-driven approaches for burst detection in water distribution systems". Urban Water Journal. 14(9). 972-983. <https://doi.org/10.1080/1573062X.2017.1279191>
- Yamijala, S., Guikema, S.D. and Brumbelow, K. (2009). Statistical models for the analysis of water distribution system pipe break data. Reliability Engineering & System Safety. 94(2). 282-293. <https://doi.org/10.1016/j.res.2008.03.011>