

## تحلیل مقایسه‌ای مدل‌های یادگیری ماشین برای پیش‌بینی شکست لوله در شبکه‌های توزیع آب

محسن ستاری<sup>۱</sup>، همایون مطیعی<sup>۲\*</sup>

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: ۱۴۰۴/۰۹/۱۹

تاریخ اصلاح: ۱۴۰۵/۰۳/۰۲

تاریخ پذیرش: ۱۴۰۵/۰۴/۰۱

### چکیده

شنا سایی و رفع سریع شکست لوله‌ها امر ضروری در مدیریت پایدار شبکه‌های توزیع آب است. در سال‌های اخیر از مدل‌های یادگیری ماشین برای پیش‌بینی شکست لوله‌ها استفاده شد. پژوهش‌های پیشین معمولاً بر کاربرد یک یا دو مدل خاص متمرکزند در حالیکه تحلیل مقایسه‌ای جامع از الگوریتم‌های مختلف و کاربرد عملیاتی آن‌ها مدنظر قرار نگرفتند. این پژوهش با هدف توسعه مقایسه عملکرد شش مدل یادگیری ماشین شامل؛ رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی و درخت تقویت شده برای پیش‌بینی شکست لوله انجام شد. مدل‌ها با استفاده از مجموعه داده‌ی شبکه توزیع آب منطقه ۲ شهر تهران که شامل ویژگی‌های فیزیکی و عملیاتی لوله‌ها با تقسیم‌بندی ۷۰ درصد برای آموزش و ۳۰ درصد برای اعتبار سنجی، توسعه یافته‌اند. نتایج بدست آمده نشان داد علی‌رغم انتظارات اولیه، مدل‌های پیچیده‌تری مانند جنگل تصادفی و درخت تقویت شده، با وجود عملکرد بی‌نقص روی داده‌های آموزشی، دچار بیش‌برازش شدید شده و برای کاربرد واقعی غیرقابل اعتمادند. مدل درخت تصمیم با مساحت زیر نمودار ۰/۹۲، حساسیت ۰/۹۳ و دقت کلی ۰/۹۲ به عنوان گزینه برتر معرفی گردید. از این مدل با توجه به تفسیرپذیری بالا نسبت به سایر مدل‌ها می‌توان در سیستم‌های مدیریت بهره‌برداری شبکه‌های توزیع آب جهت پیش‌بینی شکست لوله‌ها، استفاده کرد.

کلمات کلیدی: شبکه توزیع آب، پیش‌بینی شکست لوله، یادگیری ماشین، منحنی ROC، نرم افزار JMP Pro.

# A Comparative Analysis of Machine Learning Models for Predicting Pipe Failure in Water Distribution Networks

Mohsen Sattari<sup>1</sup>, Homayoun Motiee<sup>2\*</sup>

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: 10/12/2025

Revised: 23/05/2026

Accepted: 22/06/2026

## Abstract

Rapid identification and repair of pipe failures is essential for sustainable management of water distribution networks. Machine learning approaches have gained increasing attention in recent years for predicting pipe failures; however, most existing studies have relied on one or two specific models, lacking a comprehensive comparative assessment of diverse algorithms and their practical applicability. This study develops and systematically evaluates six machine learning models, logistic regression, decision tree, random forest, support vector machine, neural network, and gradient-boosted tree, for pipe failure prediction. The models were constructed using the dataset from the water distribution network of District 2 in Tehran, which includes detailed physical and operational pipe attributes. A 70–30 training–validation split was employed to ensure robust model development and assessment. The results reveal that, contrary to initial expectations, more sophisticated models such as random forest and gradient-boosted tree exhibited substantial overfitting, achieving near-perfect accuracy on training data but demonstrating poor generalizability to validation data, thereby limiting their suitability for operational deployment. The Decision Tree model was identified as the superior option, achieving an Area Under the Curve of 0.92, a sensitivity of 93%, and an accuracy of 92%. Given its high interpretability relative to other models, this model can be effectively utilized in the operational management systems of water distribution networks for predicting pipe failures.

**Keywords:** Water distribution network, Pipe failure prediction, Machine learning, ROC curve, JMP Pro software.

## ۱- مقدمه

بخش قابل توجهی از آب شرب شهری که با صرف هزینه‌های بسیار زیاد از مخازن تامین آب منتقل و سپس در تصفیه‌خانه‌ها تصفیه می‌شود، به دلیل هدرروی ناشی از حوادث و شکستگی لوله‌ها در شبکه‌های توزیع آب شهری تلف می‌شود (Taha et al. 2020). ماهیت مدفون و پیچیده شبکه‌های توزیع آب، بازرسی و ارزیابی وضعیت آنها را به طور قابل ملاحظه‌ای چالش‌برانگیز، نیازمند منابع و دانش فنی کرده است. با این حال، داده‌های جمع‌آوری‌شده و پردازش‌شده از طریق دیجیتالی شدن می‌توانند به

توسعه برنامه‌های مدیریتی کوتاه‌مدت و بلندمدت برای بازسازی، تعمیر و نگهداری کارآمدتر، پایدارتر و مهم‌تر از همه، چابک‌تر از این شبکه‌های آب کمک کنند. تکنیک‌های داده‌کاوی می‌توانند به تصمیم‌گیرندگان کمک کنند تا دفعات بازرسی‌های میدانی را کاهش داده و از هزینه‌های بالا، از جمله هزینه‌های تعمیرات اضطراری، هزینه‌های اختلالات گسترده در خدمات، هزینه‌های بازسازی به دلیل خسارت‌های زیست‌محیطی و ملکی و هزینه‌های بهداشت عمومی (مانند آسیب، مرگ و انتقال بیماری) ناشی از خرابی‌های فاجعه‌بار جلوگیری کنند (حسینی و ملکی، ۱۴۰۱). چنین خرابی‌هایی منجر به پیامدهای منفی اجتماعی، اقتصادی و زیست‌محیطی متعددی مانند افزایش هزینه‌های نگهداری، اختلال در ترافیک، وقفه در فرآیندهای صنعتی، به خطر افتادن امنیت عمومی و از بین رفتن پوشش گیاهی می‌شود (Omar et al. 2023). طبق گزارش‌های موجود، سالانه حدود یک میلیون حادثه در شبکه‌های توزیع آب ایران به وقوع پیوسته و بیش از ۲۱٪ از کل درآمدهای شرکت‌های آب و فاضلاب صرف ترمیم این حوادث می‌گردد (Motiee and Ghasemnejad, 2019).

شبکه‌های توزیع آب به عنوان شریان‌های حیاتی زیرساخت‌های شهری، نقشی اساسی در تأمین سلامت عمومی و توسعه اقتصادی ایفا می‌کنند. با این حال، این شبکه‌ها در سراسر جهان با چالش فرسودگی و کهنگی مواجه هستند که منجر به افزایش تعداد شکست‌های لوله می‌شود. شکست لوله‌ها نه تنها منجر به هدررفت قابل توجه آب و افزایش هزینه‌های تعمیر و نگهداری می‌شود، بلکه می‌تواند باعث قطعی خدمات، خسارت‌های جانبی و مشکلات بهداشتی گردد (مشهدی و تابش، ۱۴۰۳). در چنین شرایطی، حرکت از رویکردهای واکنشی به سمت استراتژی‌های پیشگیرانه و مبتنی بر پیش‌بینی، برای مدیریت پایدار دارایی‌های آب امری ضروری است (Karadirek et al. 2024).

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 برای سیستم‌های خط لوله جهت 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 به عنوان یک تکنیک امیدوارکننده برای پردازش و استخراج تعاملات پیچیده بین عوامل مؤثر و روندهای خرابی در سیستم‌های شبکه‌های توزیع آب ظهور کرده است (Chen et al. 2025).

مدل‌های مختلفی برای ارزیابی خط لوله، به ویژه برای پیش‌بینی شکست، توسعه داده شده‌اند. این مدل‌ها را می‌توان به مدل‌های فیزیکی، مدل‌های آماری و مدل‌های مبتنی بر یادگیری ماشین که به طور فزاینده‌ای در حال استفاده می‌باشند، طبقه‌بندی کرد. مدل‌های فیزیکی معمولاً بر فرضیات سیستم خاص متکی هستند و با محاسبه حداکثر تنش تحت نامطلوب‌ترین شرایط و مقایسه آن با مقاومت باقیمانده خط لوله، ارزیابی می‌کنند که آیا یک خط لوله شکست خواهد خورد یا خیر (Zhang and Wong, 2023). این رویکرد مستلزم دانش تخصصی و محاسبات ساختاری پیچیده است. در نتیجه، کاربرد مدل‌های فیزیکی در مواجهه با سناریوهای پیچیده و متغیر دنیای واقعی محدود است و همین امر، به کارگیری گسترده آنها را در سیستم‌های شبکه‌ای در مقیاس بزرگ، چالش‌برانگیز می‌کند. از سوی دیگر، مدل‌های آماری از قوانین آماری و قضایای مرتبط برای تطبیق رابطه بین متغیرهای مورد نظر و شکست خط لوله از طریق داده‌های تاریخی استفاده می‌کنند (Kleiner and Rajani, 2001). ساخت این مدل‌ها مبتنی بر روابط ریاضی خاص است و معمولاً فقط تعداد محدودی از متغیرها را در نظر می‌گیرد. کاربرد و اثربخشی مدل‌های آماری هنگام برخورد با عوامل مؤثر پیچیده و روابط علی پیچیده دخیل در شکست خط لوله آب محدود است (Chen et al. 2025). در مقایسه با مدل‌های آماری، یادگیری ماشین در ثبت و تحلیل روابط ظریف و غیرخطی بین متغیرها برتری دارد. اثربخشی آن در افزایش دقت پیش‌بینی‌های ترکیب خط لوله در مقایسه با مدل‌های آماری تأیید شده و با افزایش دسترسی به داده‌ها، مدل‌های یادگیری ماشین قرار است نقش فزاینده‌ای در پیش‌بینی خرابی خطوط لوله در آینده ایفا کنند (Snider and McBean, 2020).

یادگیری ماشین به دلیل تأثیر متحول‌کننده‌اش بر صنعت آب، به ویژه در ارزیابی کیفیت آب و مدیریت منابع، به طور فزاینده‌ای مورد استفاده قرار می‌گیرد. توانایی آن در تجزیه و تحلیل و پیش‌بینی داده‌ها در محیط‌های آبی متنوع مانند آب‌های سطحی، آب‌های زیرزمینی، آب آشامیدنی، فاضلاب و آب دریا بسیار ارزشمند بوده است. یادگیری ماشین در مدیریت مسائل پیچیده و غیرخطی که مدل‌های سنتی با آنها دست و پنجه نرم می‌کنند، عالی عمل می‌کند و منجر به مدل‌ها و نتیجه‌گیری‌های دقیق‌تری می‌شود. این قابلیت برای توسعه، نظارت و بهینه‌سازی سیستم‌های تصفیه و مدیریت آب بسیار مهم است و یادگیری ماشین را به ابزاری ضروری در علوم و مهندسی محیط زیست تبدیل کرده است (Zhu et al. 2022).

مدل‌های مبتنی بر یادگیری ماشین به دلیل مهارت‌شان در مدیریت مجموعه داده‌های گسترده و پیچیده، به عنوان ابزاری برجسته در حوزه پیش‌بینی شکست ظهور کرده‌اند. نقاط قوت قابل توجه آنها شامل دقت بالا، قابلیت‌های پیش‌بینی قوی، سازگاری و پتانسیل یادگیری و بهبود مداوم بر اساس داده‌های بازخورد است (Wang et al. 2022). این ویژگی‌ها منجر به پذیرش گسترده آنها در دوران اخیر شده است. برخلاف مدل‌های فیزیکی که اغلب به فرضیاتی در مورد شرایط فیزیکی سیستم شبکه متکی هستند و مکانیک اساسی را بررسی می‌کنند، مدل‌های یادگیری ماشین مسیری مبتنی بر داده را در پیش می‌گیرند. آنها بر کشف روابط بین

متغیرهای ورودی و خروجی بدون محدودیت‌های فرضیات سیستم تمرکز دارند (Solomatine and Ostfeld, 2008). این مدل‌ها با تحلیل داده‌های تاریخی مربوط به ویژگی‌های فیزیکی (مانند جنس، قطر و سن لوله) و عملیاتی (مانند فشار و عمق کارگذاری)، می‌توانند الگوهای پنهان منجر به شکست را شناسایی کرده و ریسک وقوع آن را برای هر بخش از شبکه تخمین بزنند.

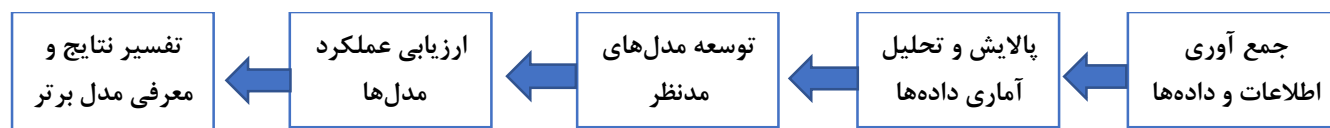
مطالعات متعددی به کاربرد الگوریتم‌هایی مانند ماشین‌های بردار پشتیبان<sup>۲</sup>، جنگل تصادفی<sup>۳</sup> و شبکه‌های عصبی<sup>۴</sup> در حوزه پیش‌بینی شکست لوله پرداخته و عملکرد قابل قبولی را گزارش کرده‌اند (Barton et al. 2022).

با این وجود، یکی از چالش‌های اصلی در این زمینه، انتخاب مناسب‌ترین مدل برای پیاده‌سازی عملیاتی است. هر الگوریتم دارای فرض‌ها، نقاط قوت و ضعف منحصربه‌فردی است. برای مثال، مدل‌های پیچیده‌تر مانند جنگل تصادفی و درخت تقویت شده<sup>۵</sup> ممکن است دقت بالایی ارائه دهند، اما در معرض خطر بیش‌برازش<sup>۶</sup> قرار دارند و تفسیرپذیری کمتری دارند. از سوی دیگر، مدل‌های ساده‌تر مانند رگرسیون لجستیک<sup>۷</sup>، به راحتی قابل تفسیر هستند اما ممکن است در مدل‌سازی روابط غیرخطی موجود در داده‌ها ناکارآمد باشند. علاوه بر این، ماهیت نامتعادل داده‌های شکست لوله، که در آن موارد شکست بسیار کمتر از عدم شکست است، ارزیابی عملکرد را پیچیده‌تر می‌کند و نیاز به معیارهای ویژه‌ای مانند منحنی مشخصه عملکرد سیستم<sup>۸</sup> (ROC) را ایجاد می‌کند (Fawcett, 2006).

در حالی که پژوهش‌های پیشین به طور کلی بر کاربرد یک یا دو مدل خاص متمرکز بوده‌اند، یک تحلیل مقایسه‌ای جامع که طیف وسیعی از الگوریتم‌ها را از ساده تا پیچیده در برابر هم قرار دهد و کاربرد عملیاتی آن‌ها را به طور دقیق بررسی کند، کمتر به چشم می‌خورد. این پژوهش با هدف پر کردن این شکاف تحقیقاتی انجام شده است. هدف اصلی این مطالعه، توسعه و ارزیابی جامع عملکرد شش مدل یادگیری ماشین شامل رگرسیون لجستیک، درخت تصمیم<sup>۹</sup>، جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی و درخت تقویت شده برای پیش‌بینی شکست لوله است. در این تحقیق جهت شناسایی بهترین مدل از نظر تعادل بین دقت، قابلیت اطمینان و کاربرد عملی برای تشخیص صحیح شکست لوله‌ها از معیارهای ارزیابی متعدد، به ویژه با تمرکز بر حساسیت و منحنی ROC، استفاده شده است.

در این پژوهش آمار شکست لوله‌های بخشی از شبکه توزیع آب در منطقه ۲ شهری تهران بین سال‌های ۱۳۹۲ تا ۱۴۰۳ جمع‌آوری و پالایش گردید. سپس با استفاده از نرم افزار JMP Pro مدل‌های پیش‌بینی شکست لوله توسعه داده شد. برای آموزش<sup>۱۰</sup> مدل‌ها از ۷۰ درصد داده‌ها و برای اعتبار سنجی<sup>۱۱</sup> از ۳۰ درصد داده‌ها استفاده گردید. در نهایت با مقایسه معیارهای ارزیابی بین مدل‌های پیش‌بینی، مدل برتر براساس تلفیقی از بالاترین حساسیت، دقت و سطح زیر نمودار منحنی ROC معرفی گشت. در شکل ۱

روندنمای پژوهش حاضر ارائه و در ادامه مراحل انجام این تحقیق شرح داده شد.



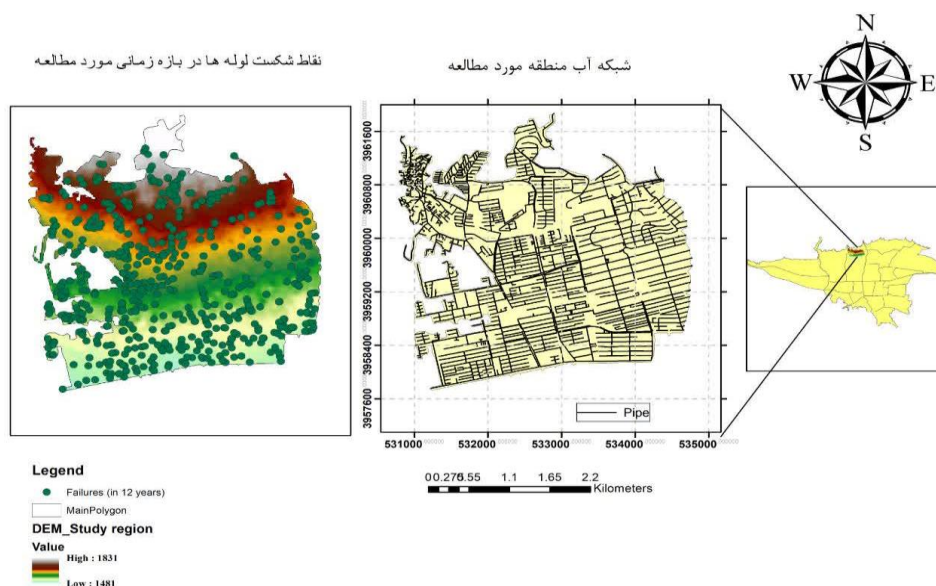
شکل ۱- روندنمای پژوهش حاضر

## ۲- مواد و روش‌ها

در این قسمت، منطقه مورد مطالعه، نرم افزار استفاده شده جهت مدل سازی، معیارهای ارزیابی عملکرد و مدل های یادگیری ماشین مدنظر، معرفی شده است.

### ۲-۱- منطقه مورد مطالعه

به منظور بررسی و ارزیابی علل حوادث رخ داده از آمار حوادث روی داده در شبکهی توزیع آب منطقه ۱ آب و فاضلاب تهران (نواحی ۱، ۸ و ۹ منطقه ۲ شهر تهران) استفاده شده است. منطقه مورد مطالعه، با ۴۷/۱ کیلومترمربع مساحت، در محدوده میانی و شمالی شهر تهران و در ارتفاع حدود ۱۴۰۰-۱۹۰۰ متر از سطح آزاد دریا قرار دارد (شکل ۲).



شکل ۲- محدوده مورد بررسی مجموعه نقاط شکست در بازه زمانی ۱۳۹۲-۱۴۰۳ (منطقه ۱ شرکت آب و فاضلاب، منطقه ۲ شهر تهران)

در این تحقیق مجموع طول خطوط لوله این نواحی برابر با ۲۰۲ کیلومتر می باشد که از ۶۴ در صد چدن نشکن<sup>۱۲</sup> (DI)، ۲۸ در صد پلی اتیلن<sup>۱۳</sup> (PE)، ۶ درصد آزیست<sup>۱۴</sup> (AC)، ۱ درصد فولادی<sup>۱۵</sup> (St) و ۱ درصد وی پی وی سی<sup>۱۶</sup> (UPVC)، تشکیل شده است.

آمار اطلاعات حوادث موضوع این پژوهش در شبکه توزیع آب و انشعابات در منطقه مورد مطالعه برای جنس لوله‌های مختلف در جدول شماره ۱ به تعداد مجموع ۵۸۲ حادثه، ارائه شده است.

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

| جنس لوله | سال ۱۳۹۲ | سال ۱۳۹۳ | سال ۱۳۹۴ | سال ۱۳۹۵ | سال ۱۳۹۶ | سال ۱۳۹۷ | سال ۱۳۹۸ | سال ۱۳۹۹ | سال ۱۴۰۰ | سال ۱۴۰۱ | سال ۱۴۰۲ | سال ۱۴۰۳ | مجموع |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-------|
| AC       | ۰        | ۰        | ۲        | ۳        | ۴        | ۳        | ۱        | ۱        | ۱        | ۲        | ۱        | ۱        | ۱۹    |
| DI       | ۲۸       | ۲۸       | ۲۷       | ۳۱       | ۲۸       | ۳۱       | ۱۳       | ۲۵       | ۲۶       | ۲۳       | ۲۹       | ۲۶       | ۳۱۵   |
| PE       | ۱۳       | ۱۸       | ۲۵       | ۲۴       | ۱۸       | ۲۶       | ۱۷       | ۱۸       | ۲۲       | ۱۶       | ۲۰       | ۲۱       | ۲۳۸   |
| ST       | ۰        | ۰        | ۰        | ۱        | ۰        | ۰        | ۱        | ۰        | ۰        | ۰        | ۱        | ۰        | ۳     |
| UPVC     | ۱        | ۰        | ۰        | ۱        | ۱        | ۰        | ۱        | ۰        | ۱        | ۰        | ۰        | ۲        | ۷     |
| مجموع    | ۴۲       | ۴۶       | ۵۴       | ۶۰       | ۵۱       | ۶۰       | ۳۳       | ۴۴       | ۵۰       | ۴۱       | ۵۱       | ۵۰       | ۵۸۲   |

## ۲-۲- نرم افزار تحلیل داده‌ها

تمامی تحلیل‌های آماری، مدل‌سازی پیش‌بینی‌کننده و ارزیابی عملکرد مدل‌ها در این پژوهش با استفاده از نرم‌افزار JMP Pro نسخه ۱۷,۲,۰ انجام شد. این نرم‌افزار یک ابزار قدرتمند در زمینه کشف آماری است که توسط شرکت SAS Institute Inc. توسعه یافته است. نرم‌افزار JMP Pro برای انجام تحلیل‌های آماری ساده و پیچیده استفاده می‌شود. این نرم‌افزار آمار را با گرافیک مرتبط می‌کند تا داده‌ها را به صورت تعاملی کاوش، درک و تجسم کند. به عنوان مثال، این ویژگی امکان ردیابی هر نقطه در یک نمودار و داده‌های مربوط به آن را فراهم می‌کند. این نرم‌افزار مجموعه‌ای جامع از ابزارهای آماری و تکنیک‌های کنترل کیفیت آماری را در یک بسته واحد ارائه می‌دهد.

این پلتفرم به دلیل ترکیب محیط کاربری گرافیکی تعاملی با الگوریتم‌های آماری و یادگیری ماشین پیشرفته، در میان محققان و مهندسان به عنوان ابزاری کارآمد برای تحلیل داده‌های پیچیده شناخته می‌شود. از آنجاییکه تمامی مدل‌های یادگیری مد نظر در این تحقیق در این نرم‌افزار وجود دارد و استفاده از آن در زمینه پیش‌بینی شکست لوله‌ها تاکنون انجام نشده است، در این مطالعه با استفاده از قابلیت‌های این نرم‌افزار، یک چارچوب تحلیلی قوی و قابل بازتولید برای توسعه، ارزیابی و مقایسه جامع شش مدل یادگیری ماشین، ایجاد شد (Gross and Harris, 2013).

متغیرهای وابسته در توسعه مدل شامل؛ طول، عمر، قطر، عمق نصب، فشار کاری و جنس لوله‌ها بوده است. در شکل ۳ نمونه اطلاعات ورودی به نرم‌افزار برای توسعه مدل‌های یادگیری ماشین، نشان داده شده است.

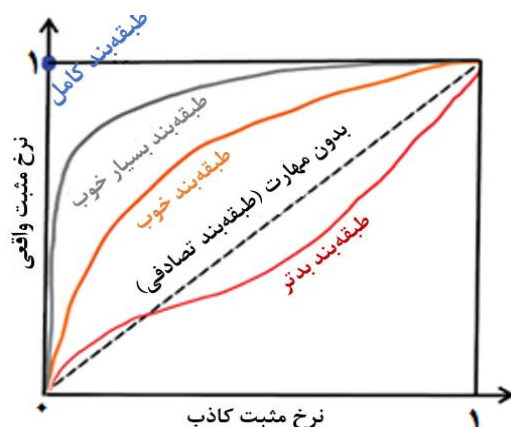
|    | Label         | Dep (m) | Dia (mm) | Material | Len (m) | Pr (mH2O) | Age (year) | Failure | Validation | Failure_Count |
|----|---------------|---------|----------|----------|---------|-----------|------------|---------|------------|---------------|
| 1  | P-277         | 1.1     | 60       | DI       | 14.3    | 38.5      | 20         | 0       | 0          | 0             |
| 2  | layelole-280  | 1.1     | 60       | DI       | 14.5    | 62.5      | 27         | 0       | 1          | 0             |
| 3  | P-128         | 1.1     | 60       | DI       | 15.2    | 36.5      | 21         | 0       | 0          | 0             |
| 4  | layelole-2407 | 1.1     | 63       | UPVC     | 13      | 69        | 29         | 1       | 0          | 1             |
| 5  | layelole-4860 | 1.3     | 63       | PE       | 13.1    | 48.5      | 21         | 0       | 1          | 0             |
| 6  | layelole-2475 | 1.1     | 63       | PE       | 13.2    | 70        | 20         | 0       | 0          | 0             |
| 7  | layelole-3968 | 1.2     | 63       | PE       | 13.8    | 45        | 21         | 0       | 0          | 0             |
| 8  | layelole-3207 | 1.1     | 60       | DI       | 42.1    | 34.5      | 21         | 0       | 0          | 0             |
| 9  | layelole-553  | 1.1     | 63       | UPVC     | 2       | 69        | 28         | 0       | 0          | 0             |
| 10 | layelole-2910 | 1.2     | 63       | PE       | 2       | 34        | 20         | 0       | 0          | 0             |
| 11 | layelole-1610 | 1.2     | 63       | UPVC     | 2.1     | 69        | 15         | 0       | 0          | 0             |
| 12 | layelole-3246 | 1.2     | 63       | PE       | 2.7     | 45        | 22         | 0       | 0          | 0             |
| 13 | layelole-4079 | 1.2     | 63       | PE       | 2.7     | 38.5      | 23         | 0       | 1          | 0             |
| 14 | layelole-4784 | 1.2     | 63       | PE       | 2.9     | 52.5      | 21         | 0       | 1          | 0             |
| 15 | layelole-1960 | 1.2     | 63       | PE       | 2.9     | 38.5      | 22         | 0       | 1          | 0             |
| 16 | layelole-4252 | 1.1     | 63       | PE       | 2.9     | 29        | 29         | 0       | 0          | 0             |
| 17 | layelole-4280 | 1.1     | 63       | PE       | 3       | 69        | 24         | 0       | 0          | 0             |
| 18 | layelole-2677 | 1.1     | 63       | PE       | 3.1     | 58        | 24         | 0       | 0          | 0             |
| 19 |               |         |          |          |         |           |            |         |            |               |

شکل ۳- نمونه اطلاعات ورودی به نرم افزار JMP Pro برای توسعه مدل های یادگیری ماشین

### ۳-۲- معیارهای ارزیابی مدل - منحنی مشخصه عملکرد سیستم (ROC):

یک روش کمی ارزیابی در مسائل طبقه بندی است که با ROC نشان داده می شود. در این روش هدف این است که توانایی آسیب شناسی یک سامانه دقیقاً ارزیابی شود. کارایی الگوریتم های دسته بندی دو دویی<sup>۱۷</sup> معمولاً به وسیله شاخص هایی مثل حساسیت<sup>۱۸</sup> و ویژگی<sup>۱۹</sup> سنجیده می شوند. اما در نمودار ROC هر دوی این شاخص ها ترکیب شده و به صورت یک منحنی نمایش داده می شوند. اغلب برای بررسی کارایی الگوریتم های دسته بندی یا ایجاد داده های رسته ای از منحنی ROC استفاده می کنند. این موضوع در شاخه یادگیری ماشین با نظارت، بیشتر مورد توجه قرار گرفته است. در زمینه تحلیل شکست لوله در شبکه های آب شهری، حساسیت، احتمال پیش بینی صحیح خرابی لوله و ویژگی، احتمال پیش بینی صحیح عدم وجود شکست لوله است (Sattari et al. 2025).

منحنی ROC با رسم نرخ مثبت واقعی<sup>۲۰</sup> (TPR) در برابر نرخ مثبت کاذب<sup>۲۱</sup> (FPR) در تنظیمات آستانه های مختلف ایجاد می شود. نرخ مثبت واقعی به عنوان حساسیت، بازایی یا احتمال تشخیص نیز شناخته می شود. نرخ مثبت کاذب نیز به عنوان احتمال هشدار کاذب شناخته می شود و می تواند به عنوان (ویژگی - ۱) محاسبه شود. در شکل ۴ منحنی ROC روابط بین نرخ مثبت حقیقی (یعنی منافع) و نرخ مثبت کاذب (یعنی هزینه) را نمایش داده شده است (Salehinasab and Goshtasbi Rad, 2025).



شکل ۴- نمودار منحنی ROC (Salehinasab and Goshtasbi Rad, 2025)

از آنجایی که TPR معادل حساسیت و FPR برابر با (ویژگی - ۱) است، نمودار ROC گاهی اوقات نمودار حساسیت در مقابل (ویژگی - ۱) نامیده می‌شود. هر نتیجه پیش‌بینی یا نمونه‌ای از یک ماتریس درهم‌ریختگی نشان‌دهنده یک نقطه در فضای ROC است.

بهترین روش پیش‌بینی ممکن نقطه‌ای را در گوشه بالا سمت چپ یا مختصات (۰ و ۱) فضای ROC به دست می‌دهد که نشان‌دهنده ۱۰۰٪ حساسیت (بدون منفی کاذب) و ۱۰۰٪ ویژگی (بدون مثبت کاذب) است. نقطه (۰ و ۱) طبقه‌بندی کامل نیز نامیده می‌شود. یک حدس تصادفی یک نقطه در امتداد یک خط مورب از پایین سمت چپ به گوشه سمت راست بالا نشان می‌دهد. قطر فضای ROC را تقسیم می‌کند. نقاط بالای مورب نتایج طبقه‌بندی خوبی را نشان می‌دهد (بهتر از تصادفی). نقاط زیر خط نشان‌دهنده نتایج بد (بدتر از تصادفی) است. خروجی یک پیش‌بینی‌کننده دائماً بد می‌تواند به سادگی معکوس شود تا یک پیش‌بینی خوب به دست آید. از منحنی ROC برای ارزیابی عملکرد معادلات رگرسیون لجستیک می‌توان استفاده کرد. به این صورت که هرچه مساحت زیر منحنی<sup>۲۲</sup> (AUC) بیشتر باشد، مدل از عملکرد مطلوب‌تری برخوردار خواهد بود (Li et al. 2019).

در مسئله پیش‌بینی دو کلاسه‌ای (کلاسه بندی دودویی) خروجی طبقه‌بندی، شامل ۴ نوع نتیجه است. به صورتی که دسته اول، همان پیش‌بینی‌های صحیح هستند که شامل مثبت صحیح<sup>۲۳</sup> (TP) و منفی صحیح<sup>۲۴</sup> (TN) میشوند و دسته دوم نیز همان پیش‌بینی‌های نادرست هستند که شامل مثبت نادرست<sup>۲۵</sup> (FP) و منفی نادرست<sup>۲۶</sup> (FN) میشوند و میتوان آنها را در قالب ماتریسی که در جدول ۲ نشان داده شده است (که به صورت ۲\*۲ می‌باشد) ارائه کرد (Wu and Liu. 2017).

جدول ۲- نتایج پیش‌بینی دو کلاسه‌ای (ماتریس درهم‌ریختگی<sup>۲۷</sup>)

| وضعیت      | مثبت پیش‌بینی شده | منفی پیش‌بینی شده |
|------------|-------------------|-------------------|
| مثبت واقعی | مثبت صحیح (TP)    | منفی نادرست (FN)  |
| منفی واقعی | مثبت نادرست (FP)  | منفی صحیح (TN)    |

برای ارزیابی عملکرد مدل‌سازی 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<sup>۲۸</sup>، حساسیت، ویژگی، دقت مثبت<sup>۲۹</sup> و F-Score از جمله معادلات هستند که برای انجام ارزیابی مدلسازی‌ها، 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).

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

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

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

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

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

## ۲-۴- مدل‌های یادگیری ماشین

در این پژوهش، شش مدل یادگیری ماشین نظارت شده توسعه داده شد. مدل‌ها به ترتیب زیر معرفی و پیاده‌سازی شدند. پیش از توسعه مدل، مجموعه داده‌های خام تحت چندین مرحله پیش‌پردازش قرار گرفت تا برای الگوریتم‌های یادگیری ماشین مورد استفاده در این پژوهش آماده شود. برای هر مدل، تمام ویژگی‌های ورودی به عنوان متغیرهای مستقل (شامل؛ طول، قطر، سن، جنس، عمق کار گذاری و فشار کاری لوله) و متغیر شکست<sup>۳۰</sup> به عنوان متغیر وابسته، در نظر گرفته شد.

### ۲-۴-۱- رگرسیون لجستیک

این مدل یک روش آماری کلاسیک برای مدل‌سازی احتمال وقوع یک رویداد دودویی است. رگرسیون لجستیک با استفاده از رابطه خطی بین متغیرهای مستقل و لگاریتم شانس وقوع شکست را برقرار می‌کند (مطیعی و قاسم نژاد، ۱۳۹۶). خروجی تولید شده از این روش بین ۰ و ۱ است که می‌تواند به عنوان احتمال شکست تفسیر شود (Rifaai et al. 2022).

- اطلاعات ورودی: تمام شش ویژگی (طول، قطر، عمق، سن، فشار و جنس).
- نتایج کلیدی برای ارزیابی: ضرایب مدل برای تفسیر تأثیر هر متغیر، و امتیاز احتمال شکست<sup>۳۱</sup> برای هر لوله که برای محاسبه معیارهای عملکرد و ترسیم منحنی ROC استفاده شد.

### ۲-۴-۲- ماشین بردار پشتیبان

مدل SVM مبتنی بر اصل کمینه‌سازی ریسک ساختاریافته است. این روش داده‌های مسئله را با استفاده از نگاشت غیرخطی به یک فضای ویژگی با ابعاد بالا نگاشت می‌کند و رگرسیون خطی را در این فضا انجام می‌دهد. این مدل یک الگوریتم قدرتمند است

که با یافتن یک ابرصفحه بهینه، داده‌ها را در فضای چندبعدی از هم جدا می‌کند. این ابرصفحه کل منطقه را به گونه‌ای تقسیم می‌کند که اشیاء با وابستگی‌های مختلف از یکدیگر جدا شوند. برای مسائل غیرخطی، از توابع کرنل استفاده می‌شود. از این توابع برای تبدیل داده‌ها به فضای ویژگی‌های بالاتر استفاده شده تا مدل بتواند مرزهای غیرخطی بین کلاس‌ها را تشخیص دهد (Shirzad et al. 2014).

- اطلاعات ورودی: تمام شش ویژگی.
- پارامترهای مدل: در این پژوهش، از کرنل خطی با پارامتر هزینه برابر با ۱ استفاده شد. این پارامتر میزان جریمه برای دسته‌بندی اشتباه را کنترل می‌کند.
- نتایج کلیدی ارزیابی: فاصله نمونه‌ها از ابرصفحه برای محاسبه امتیاز احتمال و در نتیجه ارزیابی عملکرد مدل.

#### ۲-۴-۳- شبکه عصبی مصنوعی

شبکه عصبی مصنوعی یکی از اولین مدل‌های یادگیری ماشینی بود که در مدل‌های شکست لوله به کار گرفته شد. این مدل برای بهبود دقت پیش‌بینی ضعیف مدل‌های سنتی مورد استفاده قرار گرفت و اغلب بر اساس نحوه پردازش اطلاعات توسط الگوریتم، به یک شبکه عصبی بیولوژیکی تشبیه می‌شود (Barton et al. 2022). شبکه‌های عصبی از ساختار مغز انسان الهام گرفته شده و قادر به مدل‌سازی روابط بسیار پیچیده و غیرخطی هستند. این مدل‌ها از لایه‌هایی از نورون‌های به هم متصل تشکیل شده‌اند (Kerwin et al. 2023). هر نورون درون یک لایه به هر نورون درون لایه اتصال متصل است. نورون، مجموع وزنی هر متغیر ورودی را از لایه اول تا لایه دوم (لایه پنهان) محاسبه می‌کند که از واحدهای فعال‌سازی تشکیل شده است که تابع فعال‌سازی را انجام می‌دهند و داده‌ها را به خروجی‌های غیرخطی تبدیل می‌کنند (Barton et al. 2022).

- اطلاعات ورودی: تمام شش ویژگی به عنوان نورون‌های لایه ورودی.
- ساختار مدل: از یک شبکه پیش‌خور<sup>۳۳</sup> با یک لایه پنهان و تابع فعال‌سازی تانژانت هایپربولیک (TanH) استفاده شد.
- نتایج کلیدی برای ارزیابی: امتیاز احتمال خروجی از نورون نهایی که برای طبقه‌بندی و ارزیابی عملکرد به کار رفت. این مدل به عنوان یک جعبه سیاه<sup>۳۴</sup> در نظر گرفته شد.

#### ۲-۴-۴- درخت تصمیم

این مدل یک ساختار درختی از قوانین اگر-آنگاه (if-then) ایجاد می‌کند که تصمیم‌گیری را به صریح و قابل فهم می‌سازد. هر گره درخت بر اساس یک متغیر ورودی تقسیم می‌شود تا بیشترین خلوص در گره‌های فرزند حاصل شود. این مدل از نظر تفسیر و تجسم عملکرد بسیار خوبی دارد. برای کاربرد در مورد شکست لوله، از درخت‌های تصمیم‌گیری برای شناسایی لوله‌هایی که خرابی در آنها قریب‌الوقوع است، استفاده می‌شود. مزیت اصلی درخت‌های تصمیم‌گیری، سادگی و کارایی محاسباتی این روش، هم از نظر ایجاد درخت و هم از نظر اعمال آن در تصمیم‌گیری است. (Winkler et al. 2018).

- اطلاعات ورودی: تمام شش ویژگی.
- الگوریتم: از الگوریتم درخت‌های طبقه‌بندی و رگرسیون<sup>۳۵</sup> (CART) برای ساخت درخت استفاده شد.
- نتایج کلیدی برای ارزیابی: ساختار درخت برای تفسیر قوانین تصمیم‌گیری، نمودار اهمیت متغیرها و احتمال پیش‌بینی شده در هر برگ<sup>۳۶</sup> برای ارزیابی عملکرد.

## ۲-۴-۵- درخت تقویت شده

این مدل یک روش گروهی است که به صورت متوالی تعداد زیادی درخت تصمیم می‌سازد. پس از هر تکرار تقویتی، وزن داده‌هایی که به اشتباه طبقه‌بندی شده‌اند افزایش می‌یابد و برای داده‌هایی که به درستی طبقه‌بندی شده‌اند، وزن آنها کاهش می‌یابد. هر درخت جدید، تلاش می‌کند تا خطاهای درخت‌های قبلی را از طریق تخمین‌های بهبود یافته تدریجی تصحیح کند و در نهایت یک مدل پیش‌بینی‌کننده بسیار قوی ایجاد می‌شود (Giraldo-González and Rodríguez, 2020).

- اطلاعات ورودی: تمام شش ویژگی.
- الگوریتم: از الگوریتم تقویت گرادینان<sup>۳۷</sup> استفاده شد.
- نتایج کلیدی برای ارزیابی: امتیاز احتمال تجمیعی از تمام درخت‌ها، نمودار اهمیت متغیرها و نرخ خطا بر روی مجموعه‌های آموزشی و اعتبارسنجی برای بررسی پدیده بیش‌برازش.

## ۲-۴-۶- جنگل تصادفی

جنگل تصادفی نیز یک روش گروهی دیگر است که با ساختن تعداد زیادی درخت تصمیم به صورت موازی و روی نمونه‌های تصادفی از داده‌ها، عملکرد پایداری را ارائه می‌دهد. نتیجه نهایی بر اساس آرای اکثریت درخت‌ها تعیین می‌شود. برخلاف درخت تصمیم‌گیری که در فرآیند طبقه‌بندی و پیش‌بینی فقط از یک درخت تشکیل شده است، یک جنگل تصادفی یک مکانیزم رأی‌گیری برای کلاس براساس مجموعه درختان ایجاد می‌کند که به طور قابل توجهی بر دقت مدل تأثیر می‌گذارد (Chen et al., 2022).

- اطلاعات ورودی: تمام شش ویژگی.
- الگوریتم: از الگوریتم جنگل تصادفی با ۱۰۰ درخت استفاده شد.
- نتایج کلیدی برای ارزیابی: امتیاز احتمال نهایی از آرای درخت‌ها، نمودار اهمیت متغیرها که بر اساس کاهش ناخالصی در تمام تقسیم‌ها محاسبه و عملکرد مدل بر روی داده‌های اعتبارسنجی بررسی می‌شود.

## ۲-۵- معیارهای ارزیابی عملکرد

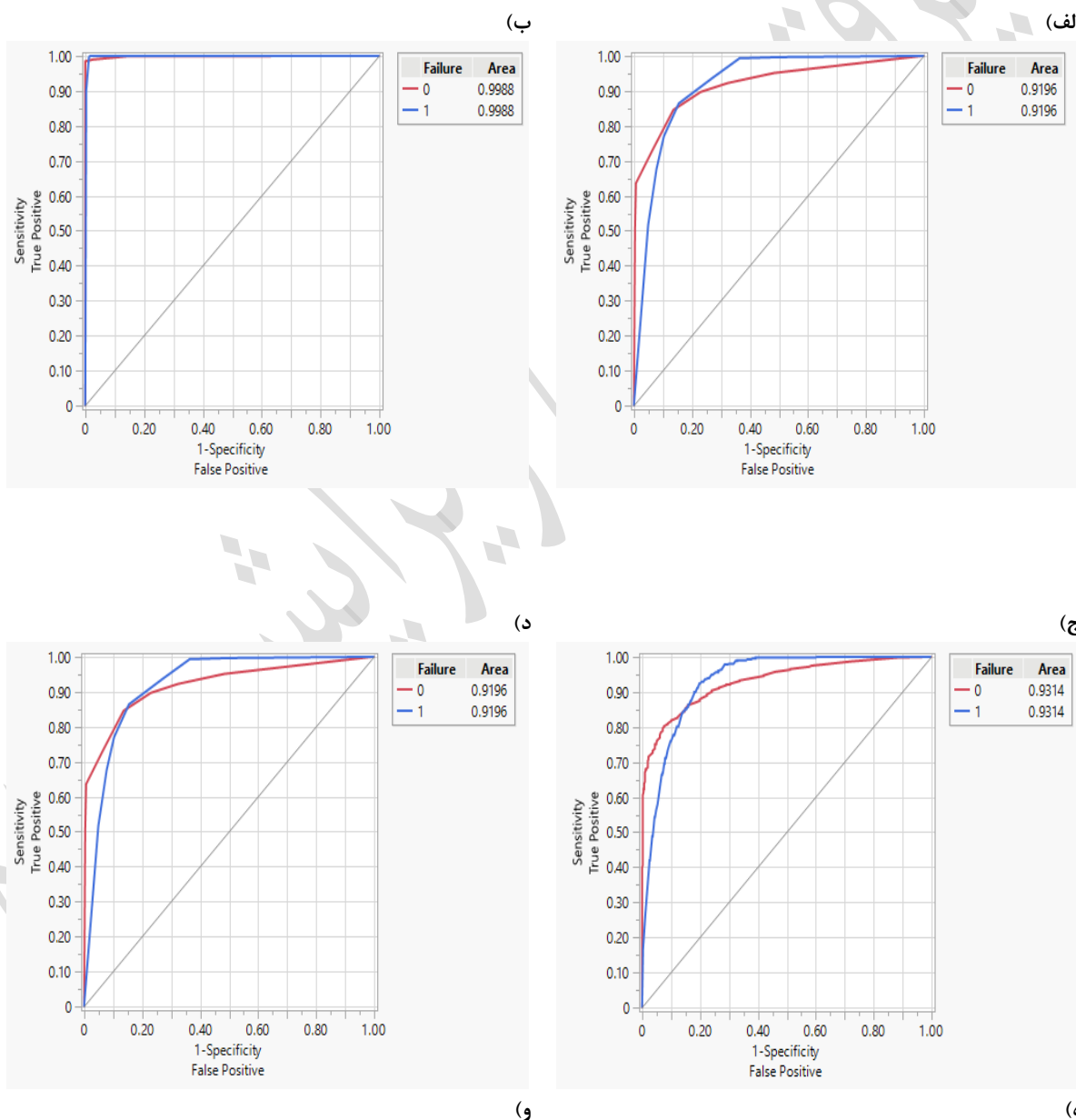
برای مقایسه منطقه‌ای مدل‌ها، از مجموعه‌ای از معیارهای ارزیابی استفاده شد. با توجه به ماهیت نامتعادل مسئله، تمرکز اصلی بر روی معیارهایی بود که عملکرد مدل در شناسایی کلاس اقلیت (شکست) را به خوبی نشان دهند.

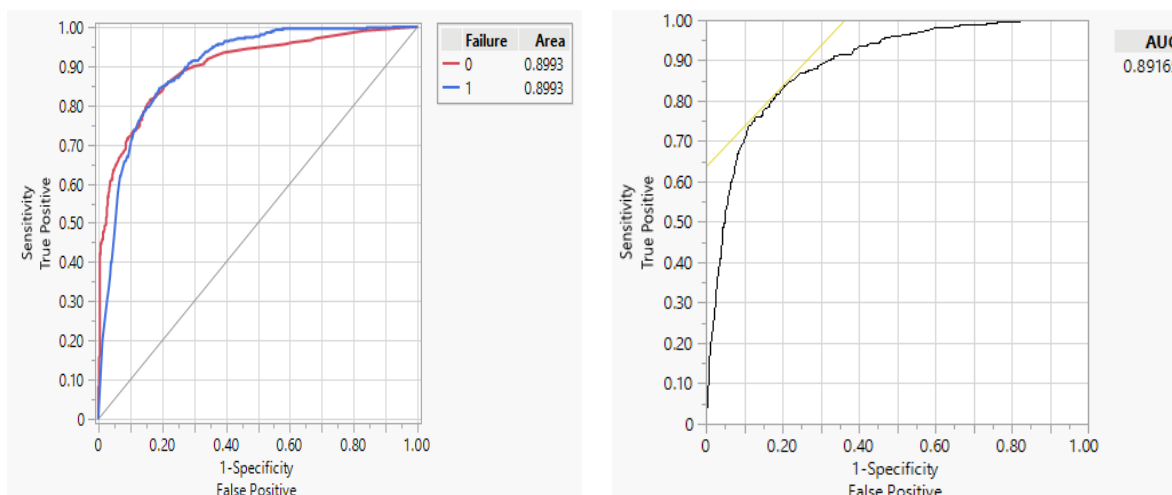
۱. ماتریس درهم‌ریختگی: برای محاسبه تعداد درست‌اشتباهی‌های هر مدل.
۲. حساسیت: توانایی مدل در شناسایی صحیح موارد شکست واقعی.

۳. ویژگی: توانایی مدل در شناسایی صحیح موارد عدم شکست واقعی.
۴. دقت مثبت: نسبت موارد پیش‌بینی شده به عنوان شکست که واقعاً شکسته‌اند.
۵. منحنی مشخصه عملکرد (ROC): نموداری که رابطه بین حساسیت و ویژگی را در تمام آستانه‌های تصمیم‌گیری نشان می‌دهد.
۶. ناحیه زیر منحنی (AUC): معیاری خلاصه‌شده برای قدرت تفکیک‌پذیری کلی مدل.

### ۳- نتایج و مقایسه مدل‌ها

در این بخش، نتایج تحلیل و مقایسه شش مدل‌ها به صورت کمی و کیفی ارائه شد. ابتدا باتوجه به توسعه مدل‌ها در نرم افزار JMP Pro نمودار ROC مربوط به خروجی مدل‌ها در شکل ۵ ارائه شده است.





شکل ۵- نمودار منحنی ROC مدل‌ها (الف) درخت تصمیم، (ب) جنگل تصادفی، (ج) شبکه عصبی، (د) درخت تقویت شده، (ه) رگرسیون لجستیک و (و) ماشین بردار پشتیبان

همچنین نتایج کلی مدل‌ها براساس معیارهای ارزیابی مد نظر در این پژوهش، به شرح جدول ۳ حاصل گردید.

جدول ۳- نتایج کلی مدل‌های پیش‌بینی شکست لوله و رتبه بندی آنها

| رتبه | مدل                 | AUC   | Accuracy | Sensitivity | Specificity | Precision | F1-Score | تحلیل اصلی                      |
|------|---------------------|-------|----------|-------------|-------------|-----------|----------|---------------------------------|
| ۱    | درخت تصمیم          | 0.92  | 92%      | 93%         | 91%         | 63%       | 75%      | تبادل برتر و قابل اعتماد        |
| ۲    | جنگل تصادفی         | 0.999 | 78%      | 100%        | 94%         | 53%       | 57%      | حساسیت خوب اما بیش‌برازش شدید   |
| ۳    | شبکه عصبی           | 0.93  | 91%      | 43%         | 97%         | 68%       | 53%      | بسیار دقیق اما حساسیت پایین     |
| ۴    | درخت تقویت شده      | 0.98  | 84%      | 99%         | 88%         | 51%       | 67%      | حساسیت عالی اما عدم پایداری     |
| ۵    | رگرسیون لجستیک      | 0.89  | 82%      | 30%         | 95%         | 62%       | 40%      | تفسیرپذیری بالا اما عملکرد ضعیف |
| ۶    | ماشین بردار پشتیبان | 0.9   | 90%      | 19%         | 99%         | 66%       | 29%      | حساسیت فاجعه‌بار                |

براساس اطلاعات شکل ۵ و جدول ۳، نتایج رتبه‌بندی مدل‌ها از نظر ملاک‌های ارزیابی و توجه به این نکته که پایداری مدل‌ها براساس کمترین اختلاف بین نتایج حاصل از داده‌های آموزشی و صحت‌سنجی می‌باشد، به شرح ذیل است.

مدل ماشین بردار پشتیبان ضعیف‌ترین عملکرد را در میان تمام مدل‌ها از خود نشان داد. حساسیت بسیار پایین (۱۸/۹۲٪) به این معناست که مدل بیش از ۸۰ درصد از شکست‌های واقعی را نادیده می‌گیرد. این عملکرد برای یک سیستم پیشگیرانه کاملاً غیرقابل قبول است و آن را به گزینه‌ای نامناسب برای این کاربرد تبدیل می‌کند. مدل رگرسیون به عنوان یک مدل پایه و قابل تفسیر، عملکرد بهتری داشت اما همچنان با محدودیت‌های جدی روبرو است. حساسیت پایین (۲۹/۷۳٪) آن را به یک گزینه غیرعملی برای پیش‌بینی فعال تبدیل می‌کند. این مدل تنها ۳۰ درصد از شکست‌های واقعی را شناسایی کرده که برای کاربرد پیشگیرانه مناسب نیست. مدل درخت تقویت شده پیچیده‌ترین و متناقض‌ترین نتایج را به همراه داشت. از یک سو، حساسیت فوق‌العاده

(۹۸/۵٪) نشان‌دهنده پتانسیل بالای آن برای شناسایی تقریباً تمام شکست‌های واقعی است. از سوی دیگر، این مدل دو نقطه ضعف دارد؛ بیش‌برازش شدید (افت عملکرد روی داده‌های اعتبارسنجی) و دقت مثبت بسیار پایین (۵۱/۲٪) که منجر به هزینه‌های عملیاتی سنگین به دلیل هشدارهای کاذب می‌شود. مدل شبکه عصبی با دقت کلی بالا (۹۱/۴۲٪) و ویژگی بی‌نظیر (۹۷/۴٪)، عملکردی بهتر از مدل‌های قبلی داشت. این مدل بالاترین دقت مثبت را در بین تمام مدل‌ها دارد، به این معنا که هشدارهای آن بسیار قابل اعتماد هستند. با این حال، این موفقیت با وجود حساسیت متوسط (۴۳/۴٪) قابلیت اتکا کمی دارد. این مدل همچنان بیش از نیمی از شکست‌های واقعی را از دست می‌دهد و آن را به یک گزینه با ریسک عملیاتی بالا تبدیل می‌کند. مدل جنگل تصادفی در داده‌های آموزشی بهترین عملکرد را نشان داد، اما با نتایج اعتبارسنجی مشخص گردید مدل دارای بیش‌برازش شدید است. افت چشمگیر عملکرد آن روی داده‌های اعتبارسنجی (حساسیت از ۱۰۰٪ به ۶۱/۲٪) نشان می‌دهد که مدل در پیکربندی فعلی خود قابل اعتماد نیست. با این حال، با حساسیت ۶۱/۲٪، همچنان بهتر از شبکه عصبی عمل می‌کند. مدل درخت تصمیم با مساحت زیر نمودار ۰/۹۲، حساسیت ۹۳٪ و دقت کلی ۹۲٪ به عنوان بهترین گزینه عملیاتی برای پیاده‌سازی داده‌ها معرفی شد. این مدل نه تنها حساسیت پایین مدل‌های خطی را جبران کرده، بلکه در تمام معیارهای کلیدی عملکرد خوبی ارائه داده است. همچنین با توجه به اختلاف کم در نتایج معیارهای ارزیابی در هر دو داده آموزشی و اعتبارسنجی، می‌توان گفت مدل دارای پایداری مطلوب و عدم بیش‌برازش بوده که این امر قابلیت اطمینان بالای مدل را نشان می‌دهد. مهم‌تر از همه، این مدل قوانین تصمیم‌گیری ساده و قابل فهمی با تفسیرپذیری بالا تولید می‌کند که به مدیران شبکه اجازه می‌دهد تا به راحتی دلایل پیش‌بینی‌ها را درک کرده و آن‌ها را در برنامه‌ریزی‌های عملی به کار گیرند. لذا با توجه به توضیحات بالا و جدول ۳ رتبه اول به مدل درخت تصمیم و رتبه ششم به مدل ماشین بردار پشتیبان اختصاص می‌یابد.

#### ۴- بحث و نتیجه‌گیری

هدف اصلی این پژوهش، توسعه و ارزیابی جامع عملکرد شش مدل یادگیری ماشین برای پیش‌بینی شکست لوله در شبکه‌های توزیع آب بود. نتایج این مطالعه نشان داد که انتخاب مدل بهینه یک فرآیند پیچیده است که به شدت به اهداف عملیاتی و اولویت‌های مدیران شبکه بستگی دارد. هیچ مدلی به تنهایی در تمام جنبه‌ها برتر نبود، اما الگوهای مشخص و قابل تفسیری از عملکرد آن‌ها شناسایی شد.

یافته‌های کلیدی حاکی از آن است که الگوریتم‌های گروهی مبتنی بر درخت به طور قابل توجهی از مدل‌های خطی و ساده‌تر عملکرد بهتری از خود نشان دادند. مدل جنگل تصادفی با دستیابی به حساسیت کامل (۱۰۰٪) و AUC نزدیک به حد نهایی

(۰/۹۹۹۸) روی داده‌های آموزشی، بالاترین پتانسیل را برای دقت پیش‌بینی از خود به نمایش گذاشت. این مدل نشان داد که با ترکیب قدرت چندین درخت تصمیم، می‌توان به یک پیش‌بینی‌کننده بسیار دقیق و قابل اعتماد رسید. با این حال، مدل درخت تصمیم به عنوان یک مدل عملگرا و متعادل، برجسته شد. این مدل نه تنها حساسیت عالی (۰/۹۲/۵۹) و دقت کلی خوبی (۰/۹۱/۵) ارائه داد، بلکه مهم‌ترین مزیت آن، تفسیرپذیری بالا و تولید قوانین تصمیم‌گیری ساده بود. این ویژگی، درخت تصمیم را به گزینه‌ای ایده‌آل برای شرکت‌های آب و فاضلاب تبدیل می‌کند که نه تنها به دنبال دقت هستند، بلکه نیاز به درک دلایل پیش‌بینی‌ها برای برنامه‌ریزی‌های عملیاتی و توجیه هزینه‌ها دارند. در مقابل، مدل‌های رگرسیون لجستیک و ماشین بردار پشتیبان به دلیل حساسیت بسیار پایین (به ترتیب ۰/۲۹/۷۳ و ۰/۱۸/۹۲) برای این کاربرد پیشنهاد نشدند. عملکرد ضعیف آن‌ها نشان داد که فرض‌های خطی و ساده‌سازی‌های آن‌ها قادر به تشخیص الگوهای پیچیده و غیرخطی منجر به شکست لوله در این مجموعه داده نیستند. مدل شبکه عصبی نیز با وجود AUC بالا، به دلیل حساسیت متوسط (۰/۴۲/۵۱) و ماهیت جعبه سیاه خود، یک رابطه ریسکی بین دقت و قابلیت فهم ارائه کرد.

نهایتاً، مدل درخت تقویت شده پتانسیل بالایی از خود نشان داد اما با مشکل بیش‌برازش شدید و دقت مثبت پایین مواجه شد. این یافته تأکید می‌کند که پیچیدگی مدل همیشه به معنای عملکرد بهتر در دنیای واقعی نیست و اعتبارسنجی دقیق برای اطمینان از پایداری مدل‌های پیچیده امری حیاتی است.

در مجموع، این پژوهش نتیجه می‌گیرد که اگرچه جنگل تصادفی از نظر تئوریک بهترین مدل است، مدل درخت تصمیم به دلیل تعادل برجسته بین دقت، تفسیرپذیری بالا، پایداری مطلوب و قوانین تصمیم‌گیری ساده و قابل فهم یک انتخاب استراتژیک و قابل استفاده برای پیاده‌سازی کاربردی در سیستم‌های مدیریت بهره‌برداری شبکه‌های توزیع آب جهت پیش‌بینی شکست لوله‌ها است.

با اینکه این پژوهش یک مقایسه تحلیلی بین شش مدل یادگیری ماشین ارائه و مدل برتر جهت پیش‌بینی شکست لوله‌ها را معرفی نموده است، لذا می‌توان با توجه به محدودیت‌های موجود، پیشنهادات زیر در مطالعات آینده در نظر گرفته شوند:

در این پژوهش از شش روش استفاده شده حال آنکه می‌توان این تعداد را با افزودن مدل‌های چون نیوی بایس<sup>۲۸</sup>، نزدیک‌ترین همسایگی<sup>۲۹</sup> و ... افزایش داد تا تحلیل جامع‌تری انجام شود. همچنین این مطالعه بر اساس داده‌های یک منطقه خاص در تهران انجام شده است. تعمیم‌پذیری نتایج به سایر شهرها با شرایط آب‌وهوایی، خاک و الگوهای مصرف متفاوت، نیاز به اعتبارسنجی بیشتر دارد. علاوه بر این، متغیرهای مهمی مانند نوع خاک، بار ترافیکی و کیفیت آب در مدل موجود نبودند. افزودن این متغیرها می‌تواند دقت مدل را در آینده افزایش دهد. در مقیاس زمانی نیز می‌توان گفت؛ مدل‌ها بر اساس داده‌های ۱۲ ساله توسعه یافتند.

برای درک کامل چرخه عمر لوله‌ها، به داده‌های طولانی‌تر نیاز است تا بتوان مدل‌ها را با دقت بیشتری توسعه داد.

## ۵- سپاسگزاری

نویسندگان کمال شکر خود را از شرکت آب و فاضلاب منطقه یک تهران و همچنین از داوران محترم که با ارائه نظرات خود باعث بهبود و ارتقای علمی این مقاله شدند ابراز می‌دارند.

نویسندگان هیچگونه تعارض منافی با هم ندارند.

## ۶- پی‌نوشت‌ها

- 1- Machine Learning
- 2- Support Vector Machine
- 3- Artificial Neural Network
- 4- Random Forest
- 5- Boosted Tree
- 6- Overfitting
- 7- Logistic Regression
- 8- Receiver Operating Characteristic
- 9- Training
- 10- Validation
- 11- Decision Tree
- 12- Ductile Iron
- 13- Polyethylene
- 14- Asbestos- Cement
- 15- Steel
- 16- Un Plasticized Polyvinyl Chloride
- 17- Binary Classifier
- 18- Sensitivity
- 19- Specificity
- 20- True Positives Rate
- 21- False Positives Rate
- 22- Area Under the Curve
- 23- True Positives
- 24- True Negative
- 25- False Positive

- 26- False Negative
- 27- Confusion Matrix
- 28- Accuracy
- 29- Precision
- 30- Failure
- 31- Probability Score
- 32- Kernel
- 33- Feedforward
- 34- Black Box
- 35- Classification and Regression Trees
- 36- Leaf Node
- 37- Gradient Boosting
- 38- Naive Bayes
- 39- K Nearest Neighbors

## ۶-مراجع

حسینی، س.ع، ملکی طولابی، ح. (۱۴۰۱). ارائه مدل برآورد نرخ شکست لوله های اصلی آب رسانی شهری با استفاده از برنامه ریزی بیان ژن (GEP). علوم و مهندسی آب و فاضلاب، ۷(۳)، ۲۸-۴۳. <https://doi.org/10.22112/jwwse.2022.307076.1288>

مشهدی تفرشی، س و تابش، م. (۱۴۰۳). محاسبه احتمال شکست کیفیت شبکه های توزیع آب به منظور برآورد ریسک با استفاده از شبکه های بیزین. علوم و مهندسی آب و فاضلاب، ۹(۱)، ۳۲-۴۲. <https://doi.org/10.22112/jwwse.2023.389324.1356>

مطیعی، ه و قاسم نژاد، س. (۱۳۹۶). کاربرد و توسعه مدل های رگرسیونی برای پیش بینی میزان شکست لوله های 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>

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, R. Chen et al. Archives of Computational Methods in Engineering, 32(6), 3821-3849. <https://doi.org/10.1007/s11831-025-10305-9>

Chen, T.Y.J., Vladeanu, G., Yazdekhashti, S. and Daly, C.M. (2022). Performance evaluation of pipe break machine learning models using datasets from multiple utilities. Journal of Infrastructure Systems, 28(2), 05022002. [https://doi.org/10.1061/\(ASCE\)IS.1943-555X.0000683](https://doi.org/10.1061/(ASCE)IS.1943-555X.0000683)

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874. <https://doi.org/10.1016/j.patrec.2005.10.010>

Giraldo-González, M.M. and Rodríguez, J.P. (2020). Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. *Water*, 12(4), 1153. <https://doi.org/10.3390/w12041153>

Gross, R.E. and Harris, S.P. (2013), July. 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>

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>

Kerwin, S., Garcia de Soto, B., Adey, B., Sampatakaki, K. and Heller, H. (2023). Combining recorded failures and expert opinion in the development of ANN pipe failure prediction models. *Sustainable and Resilient Infrastructure*, 8(1), 86-108. <https://doi.org/10.1080/23789689.2020.1787033>

Khorshidi, M. S., Nikoo, M. R., Taravatrooy, N., Sadegh, M., Al-Wardy, M., & Al-Rawas, G. A. (2020). Pressure sensor placement in water distribution networks for leak detection using a hybrid information-entropy approach. *Information Sciences*, 516, 56-71. <https://doi.org/10.1016/j.ins.2019.12.043>

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>

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>

Motiee, H., & Ghasemnejad, S. (2019). Prediction of pipe failure rate in Tehran water distribution networks by applying regression models. *Water Supply*, 19(3), 695-702. <https://doi.org/10.2166/ws.2018.137>

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.res.2021.108271>

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>

Sattari, M., Motiee, H. and Athari, M. (2025). Evaluating pipe failure risk in water distribution networks using GIS fuzzy overlay, ROC curves and machine learning logistic regression. *Water Supply*, 25(11), 1539-1555. <https://doi.org/10.2166/ws.2025.093>

Shirzad, A., Tabesh, M. and Farmani, R. (2014). A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. *KSCE Journal of Civil Engineering*, 18(4), 941-948. <https://doi.org/10.1007/s12205-014-0537-8>

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>

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)

Taha, A.W., Sharma, S., Lupoja, R., Fadhl, A.N., Haidera, M. and Kennedy, M. (2020). Assessment of water

losses in distribution networks: Methods, applications, uncertainties, and implications in intermittent supply", Resources, Conservation and Recycling, 152, 104515. <https://doi.org/10.1016/j.resconrec.2019.104515>

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)

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W. and Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees, Structure and Infrastructure Engineering, 14: 10, 1402-1411 [online]. <https://doi.org/10.1080/15732479.2018.1443145>

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>

Zhang, Y. and Wong, R.C.K. (2023). On the effect of soil water saturation on buried steel pipe responses under external load. Canadian Geotechnical Journal, 60(10), 1557-1570. <https://doi.org/10.1139/cgj-2021-0715>

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. and Ye, L. (2022). A review of the application of machine learning in water quality evaluation. Eco-Environment & Health, 1(2), 107-116. <https://doi.org/10.1016/j.eehl.2022.06.001>