

DETECÇÃO DE VAZAMENTOS EM REDES DE ABASTECIMENTO DE ÁGUA UTILIZANDO MACHINE LEARNING

Rihan Cavalcante Chalegre

Instituto Federal de Educação, Ciência e Tecnologia de Alagoas (IFAL), Campus Palmeira dos Índios

Av. Alagoas, S/N, Palmeira de Fora, Palmeira dos Índios, Alagoas
rcc2@aluno.ifal.edu.br

Mahelvson Bazilio Chaves

Instituto Federal de Educação, Ciência e Tecnologia de Alagoas (IFAL), Campus Palmeira dos Índios

Av. Alagoas, S/N, Palmeira de Fora, Palmeira dos Índios, Alagoas
mahelvson.chaves@ifal.edu.br

Fábio Rodrigo de Lima e Silva

Companhia Pernambucana de Saneamento (COMPESA)
Av. Cruz Cabugá, 1387 - Santo Amaro, Recife, Pernambuco
fabiorodrigo@compesa.com.br

Adiel Teixeira de Almeida Filho

Universidade Federal de Pernambuco (UFPE), Centro de Informática (CIn)
Av. Jorn. Aníbal Fernandes, s/n - Cidade Universitária, Recife, Pernambuco
adielfilho@cin.ufpe.br

RESUMO

Este estudo analisou a performance do algoritmo *Decision Tree* na detecção de vazamentos na rede de distribuição de água hipotética Net-1. Foram explorados experimentos com variáveis individuais e suas combinações usando a base de dados LeakDB. Análises comparativas incluíram curvas de aprendizado e validação cruzada estratificada para avaliar treinamento e teste do algoritmo, bem como evidências de *overfitting*. A performance foi medida utilizando as métricas: precisão, acurácia, recall, pontuação F1 e matriz de confusão. O experimento mais promissor foi o que combinou pressão, vazão e demanda e alcançou métricas de precisão de 96,64%, acurácia de 97,5%, recall de 100% e F1 de 98,29%. A variável demanda, destacou-se como crucial no desempenho do modelo. A implementação de machine learning no saneamento básico mostrou-se promissora, entretanto existem limitações na coleta e tratamento de dados.

PALAVRAS CHAVE. vazamentos, machine learning, perdas de água, Net-1, EPANET.

ABSTRACT

This study analyzed the performance of the Decision Tree Classifier algorithm in detecting leaks in the Net-1 network. Experiments with individual variables and their combinations were explored using the LeakDB database. Comparative analyses included learning curves and stratified cross-validation to assess performance on algorithm training and test phases for anomaly detection. No evidence of overfitting was found. Performance was measured with precision, accuracy, recall, F1 score, and confusion matrix. The most promising experiment combined pressure, flow, and demand, reached a precision of 96.64%, an accuracy of 97.5%, a recall of 100%, and an F1 of 98.29%. The demand variable stood out as crucial to the model's performance.

KEYWORDS. leaks, machine learning, water loss, Net-1, EPANET.

1. Introdução

A escassez de água é um dos grandes desafios enfrentados pela sociedade. Cerca de 97,5% da água do planeta é salgada e imprópria para consumo, e apenas 1% da água doce está disponível em rios e lagos. Essa limitação destaca a necessidade de uma gestão cuidadosa dos recursos hídricos para garantir seu uso sustentável [Lückman 2023]. O acesso universal à água potável é essencial para a qualidade de vida. Um problema recorrente nas redes de distribuição de água é a perda de água, que chega a 38,3% no Brasil, cerca de 6,5 bilhões de metros cúbicos, devido a vazamentos, erros de medição e furtos. Em termos financeiros, a perda de faturamento custou ao país R\$ 11,3 bilhões por ano [Instituto Trata Brasil 2019]. Já em 2021, essas perdas aumentaram para 7,3 bilhões de metros cúbicos, resultando em prejuízos financeiros ainda maiores [Instituto Trata Brasil 2023].

Essas perdas são agravadas por infraestruturas obsoletas e ineficiências na gestão. A recuperação de receitas por meio da redução de perdas é crucial, e as empresas de saneamento básico podem se beneficiar de soluções tecnológicas como inteligência artificial, automação e análise de dados para detectar vazamentos e otimizar a gestão do sistema [Randow Junior 2022]. Estas aplicações se tornaram mais viáveis recentemente com a disseminação de sensores acústicos, de pressão e de vazão de alta frequência, mas que sem modelos altamente preditivos e acurados não serão suficientes para uma gestão mais eficiente de perdas de água [Saiteja e Ponnappalli 2023]. As abordagens mais recentes têm empregado algoritmos tradicionais como Support Vector Machines e Isolation Forest [Glomb et al 2023], até modelos de *deep learning* [Choi e Im 2023] e *ensemble* [Ravichandran et al 2021]. Entretanto, há ainda a demanda por modelos que apresentem maior explicabilidade, lidem melhor com dados ausentes, que são comuns devido à problemas nos sensores, sejam acurados mesmo em pequenos *datasets* [Choudhary et al 2023], bem como serem capaz de responder a diversidade de situações operacionais, tais quais a ocorrência de múltiplos vazamentos ou ainda vazamentos de diferentes tamanhos [Cody et al 2020].

Neste contexto, este estudo propõe uma avaliação de performance de um modelo de aprendizado de máquina, especificamente o algoritmo Decision Tree Classifier (DTC), na detecção de vazamentos em redes de abastecimento de água, utilizando três variáveis essenciais - vazão, pressão e demanda - em diferentes combinações. A escolha do DTC deve-se à sua estrutura simples, que facilita a visualização e interpretação, permitindo identificar claramente a importância e as relações entre os atributos. Também, ele oferece boa acurácia e possui passos de aprendizado matematicamente simples, tornando-o ágil, além de não ter sido explorado no *dataset* LeakDB, escolhido por ser benchmark para estudos na área [Vrachimis et al 2018]. O LeakDB é baseado na rede de distribuição de água Net-1, uma arquitetura de Rede de Distribuição de Água (RDA) simples, pequena e virtual, frequentemente usada para testes em pequena escala. Uma implementação experimental foi realizada, replicando as condições da rede e os cenários de vazamentos, a fim de comparar o desempenho e avaliar sua aplicabilidade prática em dados hidráulicos reais.

2. Conceitos Básicos

Nesta seção, são abordados brevemente os conceitos fundamentais relacionados à vazamentos em redes de abastecimento de água, inteligência artificial, aprendizado de máquina, e os métodos específicos utilizados no estudo, como o algoritmo DTC, o método CART, curva de aprendizado e a validação cruzada. Também são descritas as métricas de avaliação de desempenho que são essenciais para medir a performance do modelo aplicado.

2.1. Vazamentos em Redes de Abastecimento de Água

Um dos aspectos importantes monitorados no setor de saneamento é a perda de água, que ocorre durante todas as operações do sistema de abastecimento de água. Em geral, as maiores perdas de água ocorrem nas redes urbanas de distribuição de água que abastecem os setores industriais e comerciais, bem como o consumo doméstico.

A definição de perda de água é que toda perda real ou aparente de água, bem como todo o consumo não autorizado que determina aumento do custo de funcionamento ou que impeça a realização plena da receita operacional [COPASA 2003].

Segundo Páscoa [2009], a IWA divide as perdas de água em: perdas reais (físicas), essas são o volume de água que não chega ao consumidor final. Essas perdas ocorrem por conta de vazamentos em redes e ramais ou, ainda, de vazamentos e extravasamentos em reservatórios. Já as perdas aparentes ou comerciais (não físicas), se referem aos volumes consumidos, mas não contabilizados, fruto de fraudes (roubo de água), falhas de cadastro, ligações clandestinas, ou na imprecisão dos equipamentos dos sistemas de micromedição.

2.2. Inteligência Artificial e Machine Learning

De acordo com Alves e Renzano [2020], a Inteligência Artificial (IA) é um campo da ciência da computação voltado para exploração de extensos bancos de dados e de algoritmos capazes de desempenharem várias funções, incluindo raciocínio, tomada de decisões, resolução de problemas, aprendizado, planejamento, entre outras características. O conceito de Aprendizado de Máquina (do inglês, *Machine Learning*) é uma das vertentes da inteligência artificial dedicada ao desenvolvimento de algoritmos e técnicas computacionais cujo propósito é aprender padrões a partir de dados históricos Goldschmidt [2010].

A representatividade dos exemplos de aprendizado influencia o desempenho dos algoritmos e a qualidade do conhecimento abstraído. Devido ao caráter indutivo do aprendizado, os algoritmos criam modelos que devem ser compatíveis tanto com os exemplos usados no treinamento quanto com novos exemplos. A capacidade de um modelo de ser compatível com novos dados é chamada de generalização; maior generalização torna o modelo mais útil e desejável em aplicações práticas [Goldschmidt 2010]. Rezende [2003] aborda que a indução, como forma de inferência lógica, permite a obtenção de conclusões genéricas a partir de um conjunto específico de exemplos. Esse processo envolve raciocínio que parte de um conceito específico para generalização, aprendendo conceitos por meio de inferências indutivas sobre exemplos apresentados. Contudo, as hipóteses geradas por essa indução podem ou não preservar a verdade.

Goldschmidt [2010] explica que no aprendizado indutivo, base dos principais algoritmos de aprendizado de máquina, o modelo de conhecimento é derivado de iterações sobre exemplos históricos, subdividindo-se em: (1) Aprendizado Supervisionado, onde os exemplos históricos incluem a informação esperada do modelo; e (2) Aprendizado Não Supervisionado, onde os algoritmos agrupam exemplos históricos com base em similaridades, organizando exemplos similares no mesmo grupo e exemplos distintos em grupos separados.

2.3. Decision Tree

Árvore de decisão ou *Decision Tree* são modelos estatísticos que empregam um método de treinamento supervisionado para classificar e prever dados. Sua construção envolve o uso de um conjunto de treinamento composto por entradas e saídas [Almeida Junior e Barros 2018].

Meira et al. [2008] oferecem uma perspectiva detalhada sobre as árvores de decisão, eles explicam que o nó raiz, localizado no topo da estrutura da árvore, é o primeiro nó. Os nós internos, incluindo o nó raiz, são nós de decisão, cada um contendo um teste sobre uma variável independente, e os resultados desses testes formam os ramos da árvore. As folhas, situadas nas extremidades da árvore, representam valores de predição para a variável dependente ou distribuições de probabilidade desses valores. Árvores de decisão são referidas como classificação ou regressão, dependendo se a variável dependente é categórica ou numérica, respectivamente.

O algoritmo de aprendizado supervisionado, divide os dados em subconjuntos com base em condições simples de decisão. É amplamente utilizado devido à sua simplicidade e capacidade de fornecer uma interpretação clara das decisões do modelo. Breiman et al. [1984] desenvolveram a técnica CART, que forma a base para muitas implementações de árvores de decisão modernas.

2.3.1. CART

Para a indução e criação da árvore de decisão, foi totalmente baseado no método CART (Classification and Regression Trees), desenvolvido por Breiman et al. [1984]. Esta abordagem trata tanto variáveis dependentes discretas (classificação) quanto contínuas (regressão) por meio da mesma tecnologia. Garcia [2003] destaca o impacto significativo desse trabalho no campo da aprendizagem automática e referência na maioria das bibliografias sobre mineração de dados. Basgalupp et al. [2010] descrevem a mineração de dados (do inglês Data Mining) como a ciência dedicada a explorar dados para identificar padrões desconhecidos.

O método CART foi selecionado pela sua simplicidade, intuitividade e eficácia em revelar relações ocultas nos dados, oferecendo alta precisão e desempenho [Minitab LLC 2021]. Silva [2012] explica que o algoritmo gera uma árvore binária navegável por perguntas simples de sim ou não, de maneira automatizada, com pouca intervenção humana. CART tende a produzir resultados superiores às técnicas estatísticas tradicionais, exceto em casos específicos com algoritmos mais complexos, onde a diferença é insignificante.

O critério Gini é utilizado para a seleção do melhor atributo como nó, avaliando a impureza de um nó. A fórmula do índice de Gini para um conjunto de dados “S” é:

$$gini(S) = 1 - \sum_{i=1}^k p(C_i | n)^2$$

Onde “ $gini(S)$ ” é o Índice de Gini para o conjunto de dados “S”; “ $p(C_i | n)$ ” é a probabilidade relativa da classe “ C_i ” em relação ao número total de registros “ n ” em “S”; e “ k ” é o número total de classes.

No caso de ser particionado “S” em dois subconjuntos “ S_1 ” e “ S_2 ”, um para cada ramificação, o índice de Gini dos dados particionados é definido por:

$$gini(S, A) = \frac{n_1}{n} gini(S_1) + \frac{n_2}{n} gini(S_2)$$

Em que “ $gini(S, A)$ ” é o Índice de Gini dos dados particionados após a divisão com base no critério “A”; “ n_1 ” e “ n_2 ” são o número de instâncias nos subconjuntos “ S_1 ” e “ S_2 ” respectivamente, após a divisão; “ n ” é o número total de instâncias no conjunto de dados original “S”; e “ $gini(S_1)$ ” e “ $gini(S_2)$ ” são os Índices de Gini dos subconjuntos “ S_1 ” e “ S_2 ” respectivamente. O algoritmo cresce a árvore até atingir a saturação, repetindo o processo de divisão para os nós filhos até que nenhum nó possa ser dividido mais, resultando em uma árvore binária de máxima pureza [Barbosa 2012].

2.4. Métricas de avaliação de desempenho

A **matriz de confusão** é a ferramenta que ilustra o desempenho de um método de classificação com contagens de verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). Para duas classes, não vazamento (0) e vazamento (1), as posições são interpretadas da seguinte forma [Silva 2018]:

		Previsto	
		Negativo	Positivo
Real	Negativo	VN, verdadeiro negativo	FP, falso positivo
	Positivo	FN, falso negativo	VP, verdadeiro positivo

Figura 1: Diagrama - matriz confusão.

A **acurácia** é a proporção de todas as predições corretas em relação ao total de amostras.

$$Acurácia = \frac{VP + VN}{VP + FN + VN + FP}$$

A **precisão** é a proporção de verdadeiros positivos em relação à soma de todos os valores detectados como positivos.

$$Precisão = \frac{VP}{VP + FP}$$

O **recall** é a habilidade do método em detectar corretamente os resultados classificados como positivos.

$$Recall = \frac{VP}{VP + FN}$$

A **pontuação F1** é a média harmônica que considera tanto a precisão quanto o recall.

$$Pontuação F1 = 2 * \frac{Precisão * Recall}{Precisão + Recall}$$

2.5. Curva de Aprendizado

Representa o erro associado a diferentes etapas do processo de aprendizagem. Schumacher [2020] explica que **curvas de aprendizado** podem avaliar o desempenho de um algoritmo ao longo do tempo ou com novos dados, ajudando a diagnosticar viés (*underfitting* ou subajuste) ou variância (*overfitting* ou sobreajuste) no aprendizado. Os cenários podem conter:

- Underfitting: Curvas de treino e validação altas e próximas, indicando ajuste inadequado;
- Overfitting: Hipótese ajusta-se excessivamente aos dados de treino, resultando em baixo erro de treino, mas alto erro de validação;
- Sem Anomalias Aparentes: Erros de treino e validação baixos e estáveis, com pequena diferença entre as curvas.

2.6. Validação Cruzada

A **validação cruzada** é uma técnica amplamente utilizada para avaliar os erros de predição de um modelo [James et al. 2013]. A técnica *k-fold* estratificada é usada para evitar distribuições desiguais da variável resposta, assegurando que cada partição contenha aproximadamente o mesmo percentual de observações de cada classe [Veloso 2022]. Isso resulta em amostras de treinamento e validação com distribuição semelhante, contribuindo para uma avaliação mais precisa do desempenho do modelo, especialmente em conjuntos de dados desbalanceados.

3. Conjunto de Dados "LeakDB" e Rede de Distribuição Net-1

O LeakDB (Leakage Diagnosis Benchmark), introduzido por Vrachimis et al. [2018], é um conjunto de dados destinado à avaliação de diagnósticos de vazamento em redes de distribuição de água. Ele abrange diversos cenários de vazamento gerados artificialmente, porém de forma realista, em diferentes redes de distribuição de água e sob variadas condições. Disponível no repositório GitHub "KIOS-Research/LeakDB", é uma ferramenta valiosa para pesquisadores que enfrentam limitações de acesso a dados reais devido à confidencialidade. O LeakDB ajuda a entender como diferentes métodos de detecção de vazamentos se comportam em cenários realistas.

A Net-1, descrita por Kammoun et al. [2022], é uma arquitetura virtual simples de uma rede de distribuição de água, utilizada para testes em pequena escala. Composta por nove nós, um reservatório e um tanque, ela foi utilizada para criar 1.000 cenários de vazamento. Os dados de pressão, vazão e demanda foram gerados a cada 30 minutos durante 1 ano, armazenados como arquivos .csv em cada um dos nós, e vazamentos ocorreram aleatoriamente em termos de data, duração, tipos, posições e tamanho.

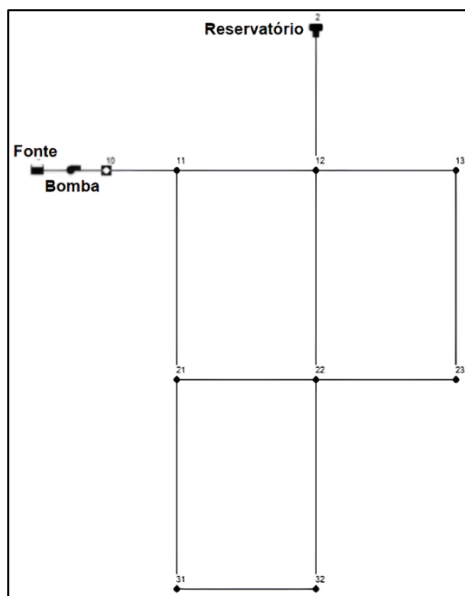


Figura 2: Arquitetura da rede de distribuição de água utilizada: Rede Net-1.

4. Tratamento de Dados e Implementação

Todo o estudo foi conduzido no ambiente Jupyter Notebook (versão 6.5.4) usando a linguagem de programação Python (versão 3.11.5). Os dados foram extraídos do LeakDB, que contém pastas organizadas por cenários. Cada cenário inclui arquivos .csv com leituras de pressão, vazão e demanda. Essas leituras foram agrupadas e unificadas em tabelas maiores, conforme ilustrado na Figura 3, que mostra o caso da variável demanda no cenário 1.

Scenario-1			
Demands			
	Index	Value	
0	1	0.0	
1	2	0.0	
2	3	0.0	
3	4	0.0	
4	5	0.0	
...	...	...	
192715	17516	-244.8	
192716	17517	-234.0	count 192720.000000
192717	17518	-226.8	mean -2.883026
192718	17519	-205.2	std 99.131152
192719	17520	-187.2	min -723.600000
			25% 0.000000
			50% 25.200000
			75% 36.000000
			max 320.400000
192720 rows x 2 columns			

Figura 3: Tabela de demandas unificada com informações referente aos 9 nós para o cenário 1 e valores estatísticos.

Seguindo o exemplo da variável demanda, foi repetido o mesmo processo para todos os cenários e extraído de cada um deles, os dados estatísticos **média**, **máximo** e **desvio padrão**, formando assim quase uma tabela que pode ser usada pelo algoritmo. Essa tabela ainda não pode ser usada por conta do efeito da magnitude de cada vazamento. Para tratar os dados e sanar esse problema, foi aplicada a técnica de padronização Z-Score, que normaliza os dados com base na média e desvio padrão extraídos dos dados estatísticos de cada uma das colunas. Isso reduz o efeito da magnitude das diferentes variáveis, permitindo uma melhor interpretação pelo algoritmo de classificação.

Cenário		D Méd.	D Máx.	D Std.	É vazamento?
0	Scenario-1	-2.883026	320.4	99.131152	1.0
1	Scenario-10	0.001868	320.4	81.687257	0.0
2	Scenario-100	0.000318	291.6	85.159941	0.0
3	Scenario-1000	-1.371164	324.0	92.865895	1.0
4	Scenario-101	-0.002522	298.8	78.900703	0.0
...	...	...	...	...	...
995	Scenario-995	-16.773232	324.0	135.530231	1.0
996	Scenario-996	-1.469328	313.2	93.650466	1.0
997	Scenario-997	-3.352304	320.4	97.624059	1.0
998	Scenario-998	-2.257528	331.2	98.418151	1.0
999	Scenario-999	0.000336	284.4	86.161431	0.0

1000 rows x 5 columns

	D Méd.	D Máx.	D Std.	É vazamento?
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	-4.075494	313.966800	98.782273	0.741000
std	6.326247	23.272105	21.009818	0.438305
min	-43.779433	262.800000	75.072454	0.000000
25%	-5.661814	302.400000	84.302154	0.000000
50%	-1.332170	309.600000	89.672218	1.000000
75%	-0.003442	324.000000	107.318682	1.000000
max	0.006949	457.200000	203.891516	1.000000

Figura 4: Tabela Demanda sem a padronização Z-Score

A padronização Z-Score é dada por:

$$z = \frac{x - \mu}{\sigma}$$

Onde “z” é o Z-Score; “x” é o valor da tabela maior que será aplicado a padronização; “μ” é a média da coluna, representada na tabela menor por “mean”; e “σ” é o desvio padrão da coluna, representado na tabela menor por “std”.

Esses valores estatísticos indicados na Figura 5, **Z-Scores de Demanda média, máxima e desvio padrão**, servem como atributos de entrada para o modelo de árvore de decisão. A última coluna da tabela indica se há ou não vazamento.

	Cenário	Z-D Méd.	Z-D Máx.	Z-D Std.	É vazamento?
0	Scenario-1	0.188495	0.276434	0.016605	1.0
1	Scenario-10	0.644515	0.276434	-0.813668	0.0
2	Scenario-100	0.644270	-0.961099	-0.648379	0.0
3	Scenario-1000	0.427478	0.431126	-0.281601	1.0
4	Scenario-101	0.643821	-0.651716	-0.946299	0.0
...	...	...	...	...	...
995	Scenario-995	-2.007152	0.431126	1.749085	1.0
996	Scenario-996	0.411961	-0.032949	-0.244258	1.0
997	Scenario-997	0.114316	0.276434	-0.055127	1.0
998	Scenario-998	0.287369	0.740509	-0.017331	1.0
999	Scenario-999	0.644273	-1.270482	-0.600712	0.0

1000 rows x 5 columns

Figura 5: Tabela Demanda com a padronização Z-Score.

Os atributos estatísticos, como média, máxima e desvio padrão, foram calculados para cada cenário, gerando as colunas "P Méd.", "P Máx.", "P Std.", "Q Méd.", "Q Máx.", "Q Std.", "D Méd.", "D Máx." e "D Std.". Esses atributos foram padronizados pelo cálculo do Z-Score, compondo o conjunto final de dados de entrada para o algoritmo DTC, enquanto a coluna "É vazamento?" indicava a presença ou ausência de vazamento.

Repetindo o procedimento para pressão e vazão, foi gerada uma tabela completa, pronta para uso pelo algoritmo. Após a padronização, o modelo DTC foi executado, com os dados divididos em 80% para treinamento e 20% para teste, garantindo a avaliação do desempenho do modelo em dados não vistos e uma melhor generalização dos resultados.

	Cenário	Z-P Méd.	Z-P Máx.	Z-P Std.	Z-Q Méd.	Z-Q Máx.	Z-Q Std.	Z-D Méd.	Z-D Máx.	Z-D Std.	É vazamento?
0	Scenario-1	0.142659	-0.216987	0.064719	-0.080975	0.154192	0.099214	0.188495	0.276434	0.016605	1.0
1	Scenario-10	0.768976	0.427001	-0.410192	-0.626473	-0.551574	-0.864870	0.644515	0.276434	-0.813668	0.0
2	Scenario-100	0.551613	-0.863836	-0.646639	-0.576007	-0.622151	-0.732205	0.644270	-0.961099	-0.648379	0.0
3	Scenario-1000	0.262945	-0.002324	-0.235550	-0.343990	-0.269267	-0.210252	0.427478	0.431126	-0.281601	1.0
4	Scenario-101	0.751263	-0.068154	-0.432762	-0.843999	-0.613328	-0.955905	0.643821	-0.651716	-0.946299	0.0
...	...	...	...	...	...	...	...	...	...	...	...
995	Scenario-995	-2.045183	1.840911	0.319363	2.367696	-0.119292	1.834946	-2.007152	0.431126	1.749085	1.0
996	Scenario-996	0.467447	0.793358	-0.084064	-0.458133	2.297957	-0.334227	0.411961	-0.032949	-0.244258	1.0
997	Scenario-997	0.130783	0.006262	-0.091236	0.106218	-0.057538	-0.040441	0.114316	0.276434	-0.055127	1.0
998	Scenario-998	0.225279	0.126473	0.032364	0.057623	-0.004605	-0.004313	0.287369	0.740509	-0.017331	1.0
999	Scenario-999	0.509852	-2.011565	-0.711974	-0.493987	-0.630973	-0.642468	0.644273	-1.270482	-0.600712	0.0

1000 rows x 11 columns

Figura 6: Tabela completa finalizada no tratamento de dados.

Os experimentos focaram em desenvolver um modelo preditivo para detectar vazamentos na rede de água, utilizando dados de pressão, vazão e demanda em diversos pontos. Foram realizados três experimentos separados para cada variável de entrada e quatro experimentos combinando diferentes variáveis. O modelo DTC foi treinado e testado para prever vazamentos. Durante um vazamento, a pressão diminui e a demanda aumenta significativamente, indicando a presença de um vazamento. A combinação das variáveis (pressão, vazão e demanda) permite uma compreensão mais completa dos padrões associados aos vazamentos, melhorando a capacidade de detecção.

5. Análises e Resultados

Para avaliar o desempenho do modelo de detecção de vazamentos na rede de água, foi conduzido os 7 experimentos utilizando 800 cenários para treinamento e 200 para teste. No total, 741 cenários continham vazamentos e 259 não continham. Devido ao desbalanceamento significativo dos dados, com 741 cenários de vazamento e 259 de não vazamentos, a validação cruzada é ainda mais crucial para garantir que o modelo não sofra de *overfitting*. Foi aplicado o método *k-fold* estratificado com "n_splits=5", dividindo o conjunto de dados em cinco partes iguais. Em cada iteração, quatro partes foram usadas para treinamento e uma para teste, repetindo o processo cinco vezes para que cada parte fosse utilizada tanto para treino quanto para teste. Utilizamos a validação cruzada estratificada com o F1-score como métrica de avaliação (scoring='f1'). Com base nos resultados dos experimentos, não há evidências claras de *overfitting*, pois os scores da validação cruzada estratificada são consistentes, sem grande disparidade entre os scores individuais e a média. Isso sugere que os modelos estão generalizando bem para os dados de teste. Após executar os 7 experimentos, o resultado das métricas de desempenho do modelo para os dados de teste são apresentados na Tabela 1.

Dados aplicados ao modelo	Acurácia	Precisão	Recall	F1
Experimento 1 - Pressão	0,825	0,861	0,903	0,881
Experimento 2 - Vazão	0,84	0,878	0,903	0,890
Experimento 3 - Demanda	0,97	0,966	0,993	0,979
Experimento 4 - Pressão e Vazão	0,815	0,859	0,889	0,874
Experimento 5 - Pressão e Demanda	0,955	0,947	0,993	0,969
Experimento 6 - Vazão e Demanda	0,965	0,954	1,000	0,976
Experimento 7 - Pressão, Vazão e Demanda	0,975	0,966	1,000	0,983

Tabela 1: Resultados para os dados de teste do modelo de classificação Árvore de Decisão.

O Experimento 1 foi o de menor desempenho dos que constam somente uma variável. No Experimento 2 houve uma melhora quase que insignificante. Já no Experimento 3, foi possível se chegar a um desempenho excelente, em relação aos outros dois, com 97% de acurácia e F1-score de 97,9%. Quando feito as combinações de duas variáveis, se observa uma melhora significativa principalmente nos experimentos que contêm a variável demanda combinada nos experimentos. Mas ainda assim, não superaram o experimento 3.

O modelo teve o melhor desempenho quando realizada a combinação de todas as variáveis: pressão, vazão e demanda. Esses resultados mostram que a inclusão de múltiplas variáveis melhora significativamente a capacidade do modelo de detectar vazamentos. Isso reforça a importância de usar uma abordagem multivariada para melhorar a performance dos sistemas de detecção de vazamentos em redes de abastecimento de água, entretanto os custos operacionais podem ser maiores, devido a demanda por múltiplos sensores.

5.1. Curvas de Aprendizado da Rede Net-1

Prosseguindo, foi analisado as curvas de aprendizado do modelo em cada um dos sete experimentos. Cada um explorou uma combinação específica de dados para detectar vazamentos na rede de abastecimento de água. O sombreamento, que mostra o intervalo de confiança, mostra a variação esperada na pontuação média do teste conforme aumentava o tamanho do conjunto de treinamento, como demonstrado na Figura 7.

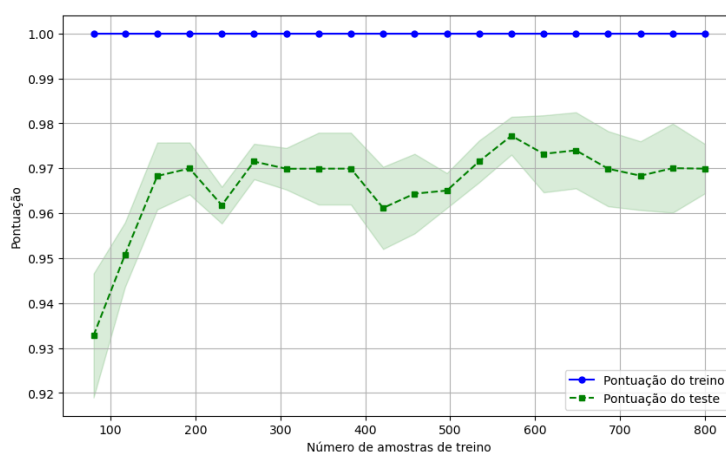


Figura 7: Curva de Aprendizado - Experimento 3 (Demanda).

As curvas de aprendizado ilustram como o modelo se ajusta aos dados de treinamento à medida que a quantidade de dados aumenta, durante treinamento e teste de cada experimento. A curva de treino indica como o modelo se ajusta aos dados de treinamento, enquanto a curva de teste mostra sua capacidade de generalizar para novos dados.

A Figura 8 mostra que os experimentos que incluíram a variável demanda tiveram um impacto significativo no desempenho. Enquanto os experimentos com uma ou duas variáveis independentes mantiveram estabilidade em suas pontuações de teste, os que incluíram demanda mostraram uma melhoria consistente.

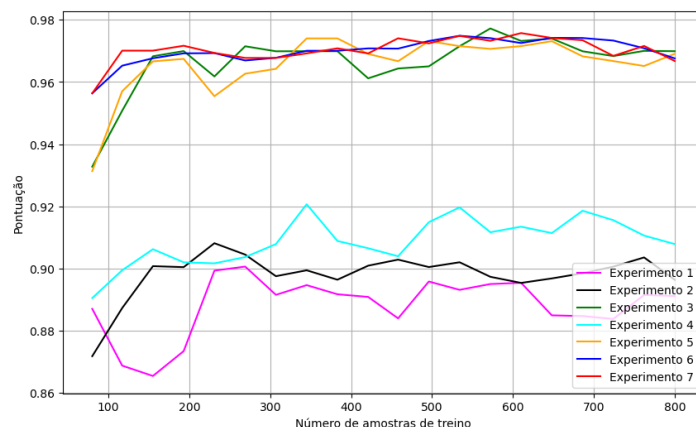


Figura 8: Curvas de Aprendizado - Pontuação Média de Teste.

Para essa análise, foi optado pela métrica F1-Score, que equilibra precisão e recall, sendo especialmente útil devido ao desbalanceamento dos dados entre vazamentos e não vazamentos. Na detecção de vazamentos de água, onde é crucial tanto identificar uma maior quantidade de vazamentos (alto recall) quanto garantir que as detecções são precisas (alta precisão), o F1-Score se destaca como a melhor métrica. Em resumo, todos os experimentos alcançaram uma pontuação de treino satisfatória, com métricas de teste estáveis, após um certo número de iterações, mostrando uma boa capacidade de generalização.

Destaca-se o Experimento 3, focado exclusivamente na variável demanda, que teve uma evolução significativa de desempenho. Já o Experimento 7 se destacou como o líder, por sua performance em geral.

6. Conclusão

Os resultados obtidos a partir da execução do modelo com um conjunto de dados contendo 1000 cenários indicam uma alta precisão e confiabilidade na detecção de vazamentos pela Rede Net-1. A melhor performance foi alcançada no Experimento 7, que utilizou dados de pressão, vazão e demanda, atingindo uma acurácia de 97,5%, F1-score de 98,29%, precisão de 96,64% e recall de 100%.

A validação cruzada com *k-fold* estratificado confirmou a robustez do modelo, não havendo evidências de overfitting. Esses resultados sugerem que o modelo pode realizar classificações com alta precisão e confiabilidade, melhorando a eficiência na distribuição de água e no tratamento de esgoto, com potencial para conservação de recursos naturais e aprimoramento do bem-estar da população.

No entanto, o estudo é limitado pela simplicidade da Rede Net-1, natureza virtual dos dados e ausência de comparação com outros algoritmos de *machine learning*. Futuras pesquisas podem explorar variáveis adicionais, como nível de reservatórios, dados acústicos, temperatura da água, velocidade de rotação do conjunto motobomba e a associação de todas as informações ao horário do dia. Redes de distribuição mais complexas ou dados reais podem validar e aprimorar o modelo, enquanto comparações entre algoritmos como *Random Forest*, *Support Vector Machines* (SVM) e *Artificial Neural Networks* (ANN) podem melhorar ainda mais os resultados, superando métodos tradicionais com maior precisão, eficiência e automação.

A implementação de técnicas de *machine learning* no setor de saneamento básico mostrou-se promissora, mas enfrenta desafios significativos, incluindo a coleta e tratamento preciso de dados, além da garantia da segurança e privacidade das informações. Mesmo assim, com o avanço das pesquisas, há um forte potencial para a conservação de recursos naturais e a melhoria do bem-estar da população através de uma aplicação prática e eficiente de *machine learning* na detecção de vazamentos.

Referências

- Almeida Junior, I. N. D., Barros, P. H. G. (2018). Análise dos Tipos de Gravação Sonora para Identificação da Doença de Parkinson. UFRJ - Universidade Federal do Rio de Janeiro. Rio de Janeiro.
- Alves, L. C. D. B.; Renzano, M. R. C. (2020). Um olhar sobre a Inteligência Artificial, a Internet das Coisas e as Tecnologias do Futuro. In *RBTI - Revista Brasileira em Tecnologia da Informação*, p. 17-23, Campinas.
- ANA – Agência Nacional de Águas e Saneamento Básico. (2023). Plano Estratégico Institucional (PEI) 2023-2026. Ministério da Integração Nacional e do Desenvolvimento Regional.
- Barbosa, J. M., Carneiro, T. G. D. S., Tavares, A. I. (2012). Métodos de Classificação por Árvores de Decisão. In: *Disciplina de Projeto e Análise de Algoritmos*. Universidade Federal de Ouro Preto, Ouro Preto.
- Basgalupp, M. P. (2010). LEGAL-Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo. São Carlos.
- Breiman, L. et al. (1984). Classification and Regression Trees. Chapman and Hall/CRC. New York.
- Cody, R. A., Dey, P., & Narasimhan, S. (2020). Linear Prediction for Leak Detection in Water Distribution Networks. *JOURNAL OF PIPELINE SYSTEMS ENGINEERING AND PRACTICE*, 11(1). [https://doi.org/10.1061/\(ASCE\)PS.1949-1204.0000415](https://doi.org/10.1061/(ASCE)PS.1949-1204.0000415)
- Choi, J., & Im, S. (2023). Application of CNN Models to Detect and Classify Leakages in Water Pipelines Using Magnitude Spectra of Vibration Sound. *APPLIED SCIENCES-BASEL*, 13(5). <https://doi.org/10.3390/app13052845>
- Choudhary, P., Botre, B. A., & Akbar, S. A. (2023). 1-D convolution neural network based leak detection, location and size estimation in smart water grid. *URBAN WATER JOURNAL*, 20(3), 341–351. <https://doi.org/10.1080/1573062X.2022.2164732>
- COPASA - Companhia de Saneamento de Minas Gerais. (2003). Programa de redução de perda de água no sistema de distribuição. Diretoria Técnica e de Meio Ambiente Superintendência de Desenvolvimento Tecnológico. Belo Horizonte.
- Garcia, S. C. (2003). O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul. Porto Alegre.
- Glomb, P., Cholewa, M., Koral, W., Madej, A., & Romaszewski, M. (2023). Detection of emergent leaks using machine learning approaches. *WATER SUPPLY*. <https://doi.org/10.2166/ws.2023.118>
- Goldschmidt, R. R. (2010). Uma Introdução à Inteligência Computacional: Fundamentos, Ferramentas e Aplicações. IST-Rio - Instituto Superior de Tecnologia do Rio de Janeiro. Rio de Janeiro.
- Instituto Trata Brasil. (2019). Perdas de água potável seriam suficientes para abastecer 30% da população brasileira por um ano. São Paulo.
- Instituto Trata Brasil. (2023). Água perdida no país poderia abastecer todas as 17,9 milhões de pessoas que residem nas favelas do Brasil por cerca de um ano e meio. São Paulo.

- James, G. et al. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. New York, p. 176-184.
- Kammoun, M.; Kammoun, A.; Abid, M. (2022). Experiments-based comparative evaluations of machine learning techniques for leak detection in water distribution systems. *Water Supply - IWA Publishing*, Sfax, v. 22, n. 1, p. 628-642.
- Lückman, A. P. (2023). Escassez de água: entenda por que esse problema também é seu. Portal do IFSC – Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina. <https://www.ifsc.edu.br/web/ifsc-verifica/w/escassez-de-agua-entenda-por-que-esse-problema-tambem-e-seu>. Acessado: 2024-12-04.
- Meira, C. A.; Rodrigues, L. H.; Moraes, S. A. (2008). Análise da Epidemia da Ferrugem do Cafeeiro com Árvore de Decisão. *Tropical Plant Pathology*, Brasília, v. 33, p. 115.
- Minitab LLC. (2021). CART®: Árvores de Classificação e Regressão. <https://www.minitab.com/pt-br/solutions/analytics/statistical-analysis-predictive-analytics/cart/>. Acessado: 2024-07-05.
- Páscoa, J. C. D. (2009). Estudos de Redução de Perdas de Água e Eficientização Energética no Setor de Saneamento. Dissertação (Mestrado em Engenharia da Energia) – Universidade Federal de Itajubá. Itajubá, p. 51.
- Ravichandran, T., Gavahi, K., Ponnambalam, K., Burtea, V., & Mousavi, S. J. (2021). Ensemble-based machine learning approach for improved leak detection in water mains. *JOURNAL OF HYDROINFORMATICS*, 23(2), 307–323. <https://doi.org/10.2166/hydro.2021.093>
- Randow Junior, R. V. (2022). Algoritmo de Identificação de Falhas Para Auxílio ao Planejamento Operacional em Sistemas de Distribuição de Água. Dissertação (Mestrado em Tecnologias Sustentáveis) – Instituto Federal do Espírito Santo. Vitória, p. 22.
- Rezende, S. O. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*. Editora Manole Ltda. Barueri, p. 90.
- Saiteja, S., & Ponnappalli, V. A. S. (2023). A review on smart water management in various domestic areas: an approach for water consumption and leakage perspectives. *INTERNATIONAL JOURNAL OF CRITICAL INFRASTRUCTURES*, 19(1), 1–16. <https://doi.org/10.1504/IJCIS.2023.129063>
- Schumacher, N. M. (2020). Estudo Comparativo de Modelos de Aprendizado de Máquina para Detecção de Email Spam. Monografia (Bacharelado em Ciência da Computação) – Universidade de Brasília. Brasília.
- Silva, A. M. M. (2012). Técnicas de Data Mining na Aquisição de Clientes para Financiamento de Crédito Direto ao Consumidor - CDC. Dissertação (Mestrado em Ciências: Estatística e Experimentação Agrônômica) – Universidade de São Paulo. Piracicaba.
- SILVA, M. C. (2018). Métodos de inteligência artificial para diagnóstico de defeitos em máquinas rotativas. UFSC - Universidade Federal de Santa Catarina. Florianópolis.
- VELOSO, L. T. (2022). Um estudo comparativo de técnicas de validação cruzada aplicadas a modelos para dados desbalanceados. Dissertação (Mestrado em Estatística) – Universidade de São Paulo. São Paulo.
- VRACHIMIS, S. G. et al. (2018). LeakDB: A benchmark dataset for leakage diagnosis in water distribution networks. University of Cyprus. Nicosia, Cyprus.