Encontro Técnico AESABESP
35º Congresso Nacional de Saneamento e Meio Ambiente

PREDIÇÃO DE VAZAMENTOS EM REDES E RAMAIS EMPREGANDO *CROSS-INDUSTRY STANDART PROCESS FOR DATA MINING* (CRISP-DM) E *MACHINE LEARNING*

Emerenciana Moreira Santos[(1)]

Graduada em Engenharia Civil, Especialista em Engenharia Sanitária pela UFMG, Analista de Saneamento e coordenadora Setor Estudos Técnicos (SET) - COPASA -MG.

Gustavo de Souza Groppo[(2)]

Graduado em Economia pela Universidade Federal de Viçosa (UFV), Especialista em Finanças (UFMG), Mestre em Economia Aplicada pela Universidade de São Paulo (ESALQ/USP) e Doutor em Saneamento pela Universidade Federal de Minas Gerais (DESA/UFMG). Analista de Desenvolvimento Tecnológico | Cientista de Dados - COPASA-MG**.**

Valter de Souza Lucas Junior[(3)]

Graduado em Engenharia Civil, Especialista em Engenharia Sanitária e Tecnologia Ambiental pela UFMG, Gerente da Unidade de Serviço de Hidrometria e Perdas (USHP) - COPASA-MG.

Endereço[(1)]: Rua Maria da Conceição Patrus n. 53 – Santa Amélia – Belo Horizonte – MG – CEP: 31555550 – Brasil – Tel: (31) 99807-1261 – e-mail: emereciana.santos@copasa.com.br

RESUMO

As perdas reais causadas por vazamentos em tubulações e ramais do sistema de abastecimento de água representam um problema para o saneamento, pois parte da água produzida pela concessionária não é faturada. Este é um tema relevante diante dos cenários de escassez hídrica e altos custos de energia elétrica, impactando negativamente a saúde financeira dos prestadores de serviço, uma vez que resultam em desperdício de recursos naturais, operacionais e financeiros. A identificação desses vazamentos nas redes é um desafio global, influenciado por diversas variáveis. Frequentemente, a seleção dos locais para ações de correção de vazamentos utiliza variáveis isoladas ou combinadas, dependendo dos dados disponibilizados pelas concessionárias. O uso conjunto desses dados, por meio de técnicas de aprendizado de máquina, pode ser um importante aliado para o processo de tomada de decisões. Nesse sentido, o presente estudo propõe a aplicação Cross-Industry Standart Process for Data Mining (CRISP-DM) e de Algoritmos de Machine Learning para investigar os fatores intervenientes ao processo de ruptura e vazamento em ramais e redes de distribuição de água, avaliando a possibilidade de direcionamento de ações para o controle de perdas de água. Os resultados obtidos indicam que, entre os métodos estudados, o SVM apresentou a maior acurácia.

PALAVRAS-CHAVE: Sustentabilidade, Meio Ambiente, Vazamentos.

INTRODUÇÃO E OBJETIVO

Os sistemas de abastecimento de água, devido à sua complexidade e características específicas, apresentam um certo nível de perdas durante a produção (Morais *et al.* 2010). Estima-se que o volume global de água não faturada (NRW) foi de 346 milhões de m^3/dia ou 126 bilhões de m^3/ano (Liembergerm e Wyatt, 2019). Ainda segundo os autores, no contexto da América Latina e região do Caribe, o Brasil é o país com maior nível de NRW.

No contexto brasileiro, a infraestrutura de abastecimento de água é antiga, o que acelera o processo de deterioração. Como resultado, surgem maiores desafios de manutenção, elevando os índices de perdas de água, especialmente devido ao descuido nas atividades de conservação (Morais *et al.*, 2010).

Faz-se importante destacar que esses vazamentos podem ser visíveis ou não. Portanto, determinar a localização desses vazamentos nas redes e ramais é um desafio para as empresas de abastecimento de água em todo mundo. Além da perda da água, os vazamentos ocasionam a perda de carga causando potencial intrusão de patógenos e riscos à saúde (Odhiambo et al., 2023). Além disso, existem algumas varáveis que contribuem para o vazamento de água nas redes, tais como: idade da rede, pressão do sistema de abastecimento, tipo de material que constitui

a rede, qualidade da implantação, danos ocasionados por terceiros (Christodoulou 2011, Aydogdu e Firat 2015, Kaushik et al., 2017).

Nesse sentido, a comunidade científica vem estudando alternativas capazes de proporcionar a redução do NRW (Morais *et al.* 2010). A maioria dos estudos estão relacionadas as perdas técnicas. Esta perda está ligada com os problemas no sistema de distribuição (vazamentos na adução e distribuição, vazamentos e transbordamentos em tanques de armazenamento e vazamentos em ramais de ligações e conexões) (Christodoulou 2011, Aydogdu e Firat 2015, Kaushik *et al.* 2017, Jang *et al.* 2018, Alizadeh *et al.* 2019, Wols *et al.* 2019, Snider e McBean 2020, Auma *et al*. 2022, Robles-Velasco *et al.* 2023, Karadirek *et al.* 2024).

O desenvolvimento de métodos puramente baseados em dados começou recentemente, em comparação com as duas famílias de métodos anteriores, a primeira baseada em modelo e a segunda baseada em um modelo misto/orientado a dados, causado pelo boom no uso de modelos calibrados. A maioria dos métodos baseados em aprendizagem se enquadra na categoria mista devido à necessidade de dados rotulados de um modelo hidráulico, cobrindo o conjunto completo de cenários possíveis. No entanto, métodos de aplicação adicionais ou filosofias de aprendizagem ajudaram a derivar abordagens de detecção puramente baseadas em dados. Assim, cada vez mais esquemas baseados em dados são desenvolvidos, reduzindo efetivamente a dependência do modelo hidráulico que nem sempre está disponível ou bem calibrado (Romero-Ben *et al*. 2023).

O estudo se justifica pelo fato de que devido a existência de vazamentos, as companhias de saneamento enfrentam altos índices de perdas de água e receita, decorrentes da falta de um gerenciamento adequado. Nesse contexto, os gestores do setor precisam adotar uma nova postura em seu processo de tomada de decisão, implementando ações que melhorem a eficiência no controle operacional do sistema, visando reduzir essas perdas (Morais *et al.*, 2010). O presente trabalho busca contribuir com a investigação de fatores intervenientes ao processo de ruptura e vazamento em ramais de redes de distribuição de água, avaliando a possibilidade de direcionamento de ações para o controle de perdas de água, empregando *machine learning* e *Cross-Industry Standard Process for Data Mining* (CRISP- DM) (Wirth e Hipp, 2000).

REFERENCIAL TEÓRICO

As concessionárias de abastecimento de água têm buscado cada vez mais soluções para reduzir o desperdício de água, principalmente em um ambiente de escassez hídrica. Muitos esforços têm sido feitos com o objetivo de promover uma melhor gestão deste recurso. A Tabela 1 apresenta um resumo dos estudos sobre falhas em redes de água, sumarizando-se variáveis e métodos empregados.

Christodoulou (2011) utilizou modelos de regressão baseados em vários fatores, como idade do tubo, material do tubo, tipo de incidente, diâmetro do tubo, e número de quebras, para predizer a taxa de falha em tubulações. O autor reitera o potencial do uso da compreensão da falha nas tubulações, permitindo o aperfeiçoamento das estratégias de reparos ou substituição de tubulações.

Aydogdu e Firat (2015) utilizaram Support Vector Machine para avaliação de fatores intervenientes e predição de falha em redes de distribuição de águas. As variáveis utilizadas foram diâmetro, comprimento e idade, além do histórico de falhas obtendo correlações entre 0,6 e 0,8. Os autores também discorrem sobre a dificuldade em obtenção de dados acurados e confiáveis o que pode afetar a performance dos modelos, umas vez que tais situações não possuem nexo com as variáveis explicativas ligadas à falha originada pelos meios inerentes à infraestrutura de abastecimento de água.

Kaushik *et al.* (2017) aplicaram um modelo de Regressão Logística para prever quebras de tubulação utilizando os dados: diâmetro, comprimento, material, idade, profundidade, quantidade de falhas prévias e dados operacionais sobre pressões estáticas de serviço máximas e mínimas. Segundo os autores existem desafios inerentes ao levantamento de dados hidráulicos de pressão e de vazão junto aos trechos de rede, pois a instrumentação para monitoramento destas grandezas encontra-se disponível normalmente apenas na entrada de Distritos de Medição e Controle (DMCs) em muitas companhias prestadoras de serviços de abastecimento de água. Outro problema levantado diz respeito ao modelo hidráulico da rede de distribuição de água que, muitas vezes, não se encontra total ou parcialmente construído e calibrado. Provavelmente é a principal razão pela qual a maioria dos trabalhos existentes não consideram fatores operacionais dinâmicos (perda de carga, velocidade, vazão e pressão) nos estudos.

Tabela 1 – Estudos sobre falhas em redes de água, sumarizando-se variáveis e métodos empregados.

Autores	Variáveis	Métodos
Christodoulou (2011)	Diâmetro, idade, material, comprimento, tipo do solo, proteção catódica do tubo	*Artificial Neural Network, Regression Model*

Aydogdu e Firat (2015)	Diâmetro, idade, comprimento, qtde de falhas	*Least Square Support Vector Machine, Artificial Neural Network, Generalized Regression Neural Network*
Kaushik *et al.* (2017)	Diâmetro, idade, material, comprimento, qtde de falhas, pressão hidráulica, velocidade profundidade do tubo	*Logistic Regression Model*
Alizadeh *et al.* (2019)	Diâmetro, idade, material, comprimento, pressão hidráulica	*Artificial Neural Network, Grasshopper Optimization Algorithm-based Support Vector Regression, Gaussian Regression*
Wols *et al.* (2019)	Diâmetro, idade, material, comprimento, qtde de falhas, pressão hidráulica, densidade de ligações	*Gradient-Boosted Tree, Random Forest, AdaBoost, Extra Tree Regression, Linear Regression*
Snider e McBean (2020)	Diâmetro, idade, comprimento, umidade do solo, proteção catódica do tubo	*XGBoost*
Auma *et al.* (2022)	Tipo de solo, densidade populacional, idade, profundidade, diâmetro, distância tubo à via	*Logistic Regression*
Gomes (2022)	Perda de carga, pressão, velocidade, válvulas redutoras de pressão, booster, reservatórios, material, idade, diâmetro, aspectos comerciais, tipo de solo, declividade do terreno, tipo de pavimento, tipo de rota, falha ramais	*AdaBoost, XGBoost, Gradient Boosting, Logistic Regression, Support Vector Machine, Random Forest, Bagged Decision Tree, k-Nearest Neighbors, Bagged k-Nearest Neighbors, Naïve Bayes*
Robles-Velasco *et al.* (2023)	Diâmetro, idade, comprimento, conexões, pressão media, flutuação pressão, Qtde previa de falhas, tempo desde a última falha	*Discriminant Analysis, Logistic Regression, Random Forest*
Karadirek *et al.* (2024)	Material, diâmetro, idade, pressão media, pressão de ruptura, Qtde falha	*Logistic Regression*

Alizadeh *et al.* (2019) avaliaram a quebra em redes de distribuição de água empregando os modelos *Support Vector Regression*, Regressão Gaussiana e Redes Neurais Artificiais. Os fatores empregados foram: material, diâmetro, idade, pressão hidráulica média e comprimento do tubo. Os autores concluem que os recursos escassos para reabilitação da infraestrutura podem ser melhor aplicados considerando a identificação das áreas com maior risco de falha por meio de modelos preditores.

Wols *et al.* (2019) avaliaram a pertinência da inclusão de variáveis climáticas aos modelos preditores de falha em tubos de redes de distribuição de água. Além dos dados relacionados à rede (material, diâmetro, comprimento e pressão de serviço), considerou-se também variáveis climáticas e ambientais, tais como precipitação, temperatura do ar, evapotranspiração, vento e tipo de solo. Os fatores mais importantes são material, temperatura, ano de instalação, diâmetro e tipo de solo, além das pressões operacionais. Este estudo utilizou dados de redes de distribuição calibradas, indicando que probabilidade de falha pode ser reduzida conforme o decréscimo da amplitude entre pressões diurnas e noturnas.

Snider e McBean (2020) investigaram a aplicação de modelos preditores de falha em tubulações de água, avaliando a acurácia das predições em ambientes de dados limitados e com poucas variáveis de entrada, indicando que análises baseadas apenas em idade possuem acurácia baixa. Os autores concluem que há necessidade de componentes representativos sobre os fatores intervenientes à falha, embora muitos prestadores de serviço de água utilizem análises baseadas apenas em idade por meio de ranqueamento. Em cenários compostos por poucos dados, os autores concluem que o *XGBoost* performou agregando valor à predição de falhas.

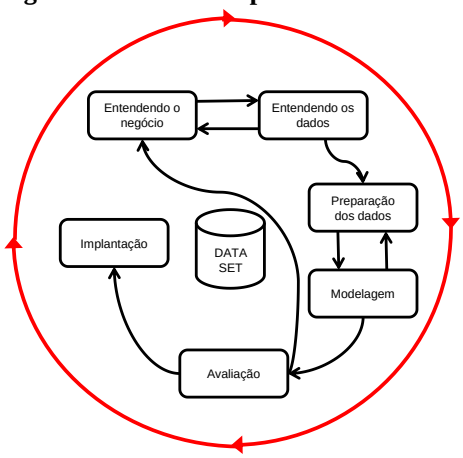
Gomes (2022) empregaram um conjunto de métodos de aprendizagem de máquina (*AdaBoost, XGBoost, Gradient Boosting, Logistic Regression, Support Vector Machine, Random Forest, Bagged Decision Tree, k-Nearest Neighbors, Bagged k-Nearest Neighbors* e *Naïve Bayes*) para prever falhas em ramais de rede de distribuição de água para a Companhia de Saneamento Ambiental do Distrito Federal (Caesb). Os resultados obtidos indicaram o modelo Adaboost como melhor método para classificação dos dados, performando acurácia de 59,23% utilizando as seguintes variáveis físicas, operacionais, ambientais e comerciais disponíveis (perda de carga, pressão, velocidade, válvulas redutoras de pressão, booster, reservatórios, material, idade, diâmetro, aspectos comerciais, tipo de solo, declividade do terreno, tipo de pavimento, tipo de rota, falha ramais). Visando

avaliar o melhor método aplicou-se as métricas Acurácia, Matriz de Confusão e importância das variáveis. Os resultados obtidos indicaram o modelo Adaboost como melhor método para classificação dos dados, performando acurácia de 59,23%. Para ganho final e adicional de performance, o AdaBoost foi submetido à hiper-parametrização, o que resultou em acurácia final de 59,70%.

MATERIAL E MÉTODOS

O *Cross-Industry Standard Process for Data Mining* (CRISP- DM) é uma metodologia de mineração de dados padrão da indústria (Wirth e Hipp, 2000) sendo a metodologia analítica mais amplamente utilizada no desenvolvimento de projetos de descoberta de conhecimento (Martínez-Plumed *et al*., 2021). O modelo CRISP-DM consiste na compreensão de negócios, na compreensão de dados, na preparação de dados, na construção, na avaliação e na implantação do modelo. Essa é uma metodologia capaz de transformar os dados da empresa em conhecimento e informações de gerenciamento. A Figura 1 mostra as seis fases do modelo de processo CRISP-DM e suas interações.

Figura 1 – Modelo de processo CRISP-DM



Fonte: Adaptado Wirth e Hipp (2000).

Qualquer projeto de mineração de dados (DM) começa com a definição da meta do projeto (objetivo) que está incluída na compreensão do negócio. Na fase de compreensão dos dados as hipóteses de informações ocultas relacionadas ao objetivo são formadas com base na experiência e em suposições qualificadas. No caso desse estudo, que tem como objetivo a detecção de consumo não autorizado, um conceito válido seria procurar novas fraudes. Na fase de preparação os dados relevantes são coletados e pré-processados. Além disso, os conjuntos de dados serão rotulados de acordo com o conhecimento de domínio especializado, de modo que as fraudes sejam representadas no conjunto de dados. Na fase de modelagem, um fluxo de trabalho de DM é construído para encontrar as configurações desejadas de parâmetros para os algoritmos selecionados e para executar a tarefa de mineração de dados nos dados pré-processados. Na fase de avaliação o modelo treinado é testado em relação a um conjunto de dados reais e os resultados da DM são avaliados de acordo com o objetivo de negócio. O conjunto de teste segue as etapas desenvolvidas nas fases de preparação e de modelagem excluindo a etapa de rotulagem. Após a avaliação do modelo estudado, tem-se a sexta, e última fase, que é pôr em produção (implantação).

Utilizou-se uma série de dados que foram coletados entre o período de 12/02/2023 a 12/04/2023 para uma área da Região Metropolitana de Belo Horizonte (RMBH), contemplando 2708 registros de falha em redes e ramais de distribuição. As variáveis empregadas são: o logradouro, o trecho da rede, o comprimento da rede, a idade da rede, a localização do tubo (via, passeio ou padrão), o sistema (sistema produtor), o sub sistema (reservatórios), o material (a classe do material, o diâmetro do tubo e a profundidade), as O.S de vazamentos de água na rua/passeio (148) e de vazamentos no padrão (149), além dos dados de temperatura e de pluviometria obtidos no site do Instituto Nacional de Meteorologia (INMET). Este conjunto de dados foram divididos em 70% para treinar e 30% para testar os modelos e garantir uma melhor avaliação dos modelos finais.

As principais ferramentas utilizadas neste trabalho estão implementadas no software R. Os métodos *Support Vector Machines with Radial Basis Function Kernel* (SVM) - (method = 'svmRadial'); *Random Forest* (RF) - (method = 'rf'); *Stochastic Gradient Boosting* (SGB) - (method = 'gbm') e *AdaBoost* (method = 'AdaBoost.M1') foram aplicados utilizando o pacote Caret (Kuhn, 2008). O modelo *Extreme Gradient Boosting* (XGBoost*)* foi implementado usando o pacote xgboost (Yuan, 2024), que é baseado no estudo de Chen e Guestrin (2016).

Geralmente os algoritmos de *machine learning* empregam a precisão como um índice de avaliação. No entanto, para um conjunto de dados desbalanceado, a avaliação dos classificadores tende a reduzir o seu efeito de classificação. Segundo He e Garcia (2009), uma maneira eficaz de avaliar um classificador no contexto desbalanceado é utilizar as taxas de erros e acertos cometidos para cada classe, sendo obtidas a partir da matriz de confusão. Portanto, múltiplas métricas devem ser empregadas objetivando medir o desempenho do algoritmo. As métricas utilizadas no presente estudo (acurácia, sensibilidade/recall, precisão, F1 score) são calculadas a partir da matriz de confusão (Figura 2), onde as classificações corretas estão registradas nas células da diagonal principal, e as classificações incorretas nas demais células. Um resultado é chamado de verdadeiro positivo (VP) quando o algoritmo classifica como positivo um caso que de fato é positivo e de verdadeiro negativo (VN), quando classifica como negativo um caso que de fato é negativo. Um resultado de falso positivo (FP), erro tipo I, é aquele em que o algoritmo classifica como positivo um caso que na verdade é negativo e de falso negativo (FN), erro tipo II, o oposto.

Figura 2 – Matriz de confusão para duas classes

		Classe prevista (prevista pelo teste)		
		Positivo	Negativo	
Classe verdadeira (confirmada pela análise)	Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)	VP + FN = Total amostra positiva
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)	FP + VN = Total amostra negativa
		VP + FP = Total previsto como positivo	FN + VN = Total previsto como negativo	

Segue uma breve apresentação dos conceitos das métricas escolhidas para este estudo, e suas respectivas fórmulas:

Acurácia: Proporção de classificações corretas, tanto de casos positivos quanto negativos. Em caso de dados desbalanceados esta medida pode induzir a uma conclusão errônea quanto ao desempenho do algoritmo empregado visto que a classe majoritária encobrirá a falha frente à classe minoritária.

$$\text{Acurácia} = \frac{VP+VN}{(VP+FN)+(FP+VN)} \quad \text{equação (1)}$$

Sensibilidade/Recall: É a proporção de verdadeiros positivos que estão identificadas corretamente.

$$\text{Recall} = \frac{VP}{VP+FN} \quad \text{equação (2)}$$

Precisão: É a proporção de verdadeiros positivos que estão realmente corretas.

$$\text{Precisão} = \frac{VP}{VP+FP} \quad \text{equação (3)}$$

F1 Score: A pontuação F1 é a média harmônica entre a precisão e a sensibilidade/recall

$$\text{F1 Score} = 2 * \frac{\frac{TP}{TP+FP} * \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \quad \text{equação (4)}$$

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A Figura 3 apresenta a importância das variáveis na predição para cada um dos modelos avaliados. Para o modelo SVM, que obteve a maior acurácia entre os modelos empregados, as variáveis que tiveram o maior peso na predição foram idade, temperatura, precipitação, trecho da rede e localização do tubo. Já os outros quatro modelos avaliados tiveram as mesmas cinco variáveis entre as mais importantes, quais são: logradouro, trecho de rede, comprimento da rede, idade da rede e temperatura. Para o modelo que apresentou a melhor performance no tocante as métricas precisão, recall e F1 Score (XGBoost) a variável que teve a maior importância foi o trecho da rede, seguidos pelo logradouro, pelo comprimento, pela idade e pela temperatura, respectivamente.

Os resultados obtidos na predição de falhas dos cinco modelos avaliados estão apresentados na Tabela 2. O SVM apresentou a maior acurácia (61,228) entre os métodos utilizados, contudo apresentou os piores resultados para as métricas recall e F1. Para todas as métricas Precisão, Recall e F1 Score, o método XGBoost obteve os melhores resultados, com valores 62,531, 67,507 e 64,923, respectivamente.

Tabela 2 – Performance da previsão dos modelos avaliados

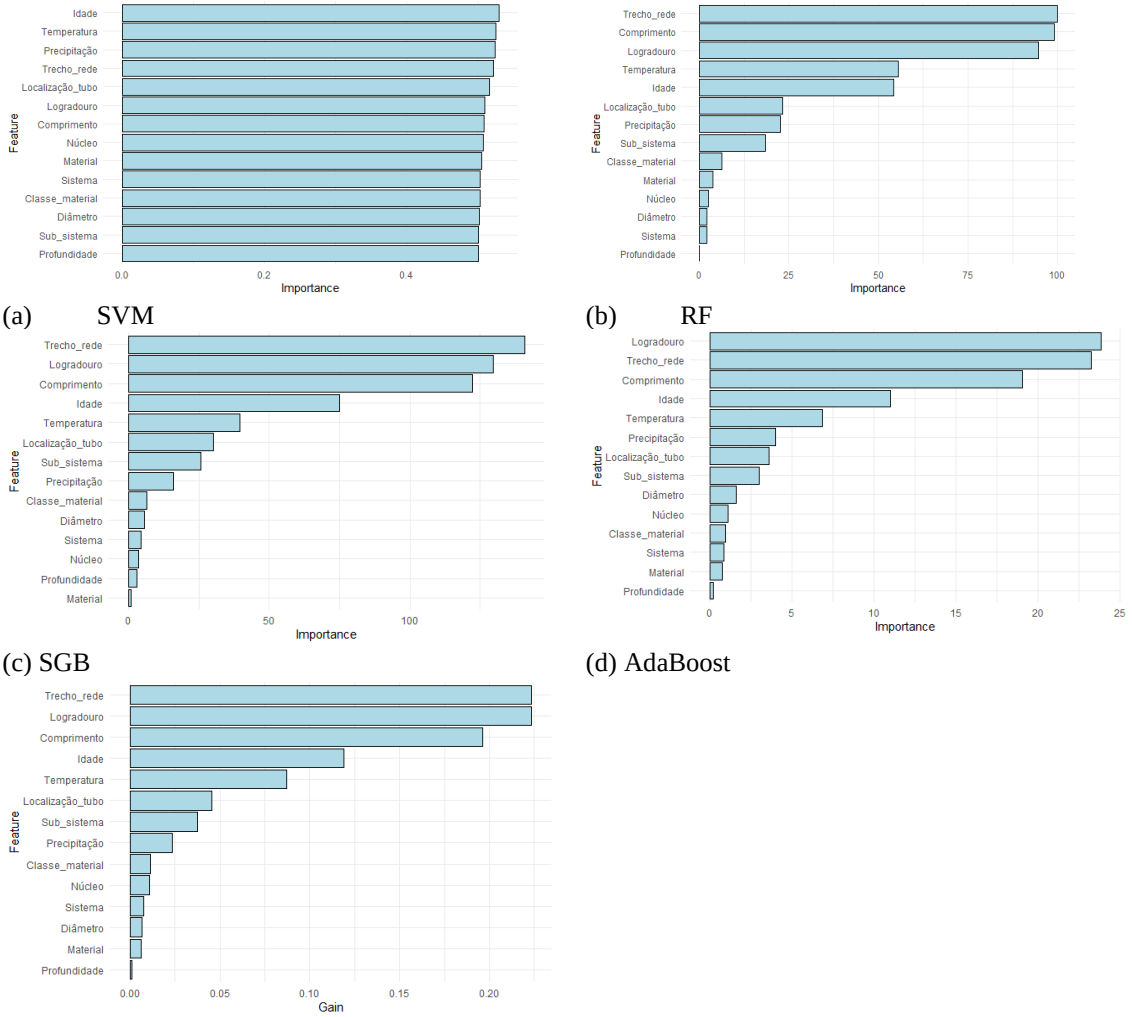
	Acurácia	Precisão	Recall	F1
SVM	**61,228**	54,237	6,349	11,368
RF	54,701	43,034	48,413	45,565
SGB	60,917	50,146	34,127	40,614

AdaBoost	60,218	49,029	40,079	44,105
XGBoost	57,919	**62,531**	**67,507**	**64,923**

Fonte: elaboração própria.

Em estudo semelhante, realizado para prever vazamentos em ramais empregando um conjunto de dados da CAESB, o modelo que apresentou a maior acurácia foi o AdaBoost, seguido pelo XGBoost com 59,23 e 57,88, respectivamente (Gomes, 2022). Neste estudo foi empregado um conjunto de 38 variáveis. Sem utilizar as variáveis operacionais (pressão), a variável inclinação foi a variável com maior importância na modelagem. Ao empregar dados de pressão, a variável de obteve maior importância foi idade, seguida pela pressão e pela inclinação.

Figura 3 – Importância das variáveis na previsão dos modelos avaliados



(a) SVM (b) RF

(c) SGB (d) AdaBoost

(e) XGBoost

Fonte: elaboração própria.

CONCLUSÕES E RECOMENDAÇÕES

Métodos de análise de dados para subsidiar o processo de compreensão e predição sobre falha em tubulações de sistemas de abastecimento de água tem sido amplamente aplicados nos últimos anos (Christodoulou 2011, Aydogdu e Firat 2015, Kaushik *et al.* 2017, Jang *et al.* 2018, Alizadeh *et al.* 2019, Wols *et al.* 2019, Snider e McBean 2020, Auma *et al.* 2022, Robles-Velasco *et al.* 2023, Romero-Ben *et al.* 2023, Karadirek *et al.* 2024), apresentando resultados relevantes para contribuição à tomada de decisão sobre o planejamento de investimentos e a gestão da infraestrutura.

Os resultados obtidos indicam que o SVM foi o método que teve a maior acurácia entre os métodos estudados. Contudo, foi o método que apresentou os piores resultados no tocante a recall e F1 Score. O método que apresentou os melhores resultados para as métricas precisão, recall e F1 Score foi o XGBoost.

Como limitação do estudo, tem-se a necessidade de ampliação da amostra e sistematização em bancos de dados. À medida que mais dados sejam coletados e sistematizados, há a tendência de que tais modelos possam evidenciar melhores performances. Adicionalmente, os resultados obtidos podem ser melhorados incluindo a variável inclinação, variável que obteve maior importância dentre as variaveis empregadas no estudo realizado por Gomes (2022) para a Caesb, quando não incluída as variáveis operacionais de pressão. Após a inclusão destas variáveis a inclinação foi a terceira variável com maior importância na predição.

REFERÊNCIAS BIBIOGRÁFICAS

1. Alizadeh, Z., Yazdi, J., Mohammadiun, S., Hewage, K., Sadiq, R. *Evaluation of data driven models for pipe burst prediction in urban water distribution systems. Urban Water Journal*, 16(2), 136-145, 2019.

2. Auma, R., Musaazi, I.G., Tumutungire, M.D., Sempewo, J.I. *Logistic pipe failure prediction models for an urban water distribution network in the developing world: a case study of Kampala water, Uganda. Water Practice & Technology*, 18(1), 264, 2022.

3. Aydogdu, M., Firat, M. *Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods. Water Resource Management*, 29, 1575-1590, 2015.

4. Brasil - Ministério do Desenvolvimento Regional. Secretaria Nacional de Saneamento – SNS. Sistema Nacional de Informações sobre Saneamento: 25º Diagnóstico dos Serviços de Água e Esgotos – 2019. Brasília: SNS/MDR, 2020.

5. Chen, T., Guestrin, C. XGBoost: *A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794, 2016.

6. Christodoulou, S. E. *Water network assessment and reliability analysis by use of survival analysis. Water Resources Management*, 25(4), 1229-1238, 2011.

7. Farley, M., Wyeth, G., Ghazali, Z., Istandar, A., Singh, S. The Manager's Non-Revenue Water Handbook: A Guide to Understanding Water Losses, United States Agency for International Developing and Ranhill Utilities Berhad: Bangkok, Thailand, Kuala Lumpur, Malaysia, 2008.

8. Gomes, C.G.N. Técnicas de Aprendizado de Máquina aplicadas à predição de vazamentos em ramais de redes de distribuição de água. Dissertação de Mestrado em Tecnologia Ambiental e Recursos Hídricos, Publicação PTARH.DM-241/2022, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 2022.

9. He, H., Garcia, E.A. *Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering*, 21, 1263-1284, 2009.

10. Jang, D., Park, H., Choi, G. *Estimation of leakage ratio using principal component analysis and artificial neural network in water distribution systems. Sustainability*, 10(3), 1-13, 2018.

11. Karadirek, I.E., Kaya-Basar, E., Akdenizc, T. A study on pipe failure analysis in water distribution systems using logistic regression. Water Supply, 24(1), 176-186, 2024.

12. Kaushik, G., Manimaran, A., Vasan, A., Sarangan, V., Sivasubramaniam, A. *Cracks under pressure? Burst prediction in water networks using dynamic metrics. 29th AAAI Conf. on Innovative Applications, Association for the Advancement of Artificial Intelligence*, San Francisco, USA, 2017.

13. Kuhn, M. *Building Predictive Models in R Using the caret Package. Journal of Statistical Software*, 28(5), 1-26, 2008.

14. Liemberger, R., Wyatt, A. *Quantifying the global non-revenue water problem. Water Supply*, 19(3), 831-837, 2019.

15. Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M.J., Flac, P. *CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33, 8, 3048-3061, 2021.*

16. Morais, D.C.; Cavalcante, C.A.V.; Almeida, A.T. *Priorização de áreas de controle de perdas em redes de distribuição de água. Pesquisa Operacional*, 30(1), 15-32. 2010.

17. Odhiambo, M., Viñas, V., Sokolova, E., Pettersson. T.J.R. *Health risks due to intrusion into the drinking water distribution network: hydraulic modelling and quantitative microbial risk assessment. Environmental Science: Water Research & Technology*, 9 (6), 1701–1716, 2023.

18. Robles-Velasco, A., Cortés, P., Muñuzuri, J., De Baets, B. *Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. Expert Systems With Applications*, 213, 119050, 2023.

19. Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V., Duviella, E. *Leak detection and localization in water distribution networks: Review and perspective. Annual Reviews in Control*, 55, 392–419. 2023.

20. Snider, B., McBean, E.A. *Watermain breaks and data: the intricate relationship between data availability and accuracy of predictions. Urban Water Journal*, 1744-9006, 2020.

21. Wirth, R., Hipp, J. *CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag: London, UK, 2000.

22. Wols, B.A., Vogelaar, A., Moerman, A., Raterman, B. *Effects of Weather Conditions on drinking Water Distribution Pipe Failures in the Netherlands. Water Supply*, 19(2), 404-416, 2019.

23. Yuan, J. *Extreme Gradient Boosting*. R package version 1.7.7.1., 2024. https://cran.r-project.org/web/packages/xgboost/xgboost.pdf