Encontro Técnico AESABESP
32º Congresso Nacional de Saneamento e Meio Ambiente

APLICAÇÃO DO MÉTODO SVM PARA IDENTIFICAÇÃO DE NÚMERO ELEVADO DE VAZAMENTOS A PARTIR DE VARIÁVEIS OPERACIONAIS EM UM SETOR DE ABASTECIMENTO

Renato Von Randow Junior[(1)]
Bacharel em Engenharia Civil pela Universidade Federal do Espírito Santo - UFES (2007), Mestrando do Programa de Tecnologias Sustentáveis do Instituto Federal do Espírito Santo (IFES – 2021). Engenheiro Civil na CESAN (Companhia Espírito Santense de Saneamento).
Reginaldo Barbosa Nunes[(2)]
Bacharel em Engenharia Elétrica pela Universidade Federal do Espírito Santo - UFES (1988), Mestre em Informática (UFES - 2003) e Doutor em Engenharia Elétrica (UFES - 2016). Professor do Instituto Federal do Espírito Santo (IFES).
Rodrigo Varejão Andreão[(3)]
Bacharel em Engenharia Elétrica pela Universidade Federal do Espírito Santo - UFES (1998), Mestre em Engenharia Elétrica pela Universidade Estadual de Campinas (UNICAMP - 2000), Doutor em Optimisation et Sûreté des Systèmes pelo Institut National Des Télécommunications (2004), França, e Pós-doutor em Processamento de Sinais Biológicos (UFES-2006). Professor do Instituto Federal do Espírito Santo (IFES).

Endereço[(1)]: **Rua Ricardo Pasolini, 63, - Centro – Santa Teresa - ES - CEP: 29650-000 - Brasil - Tel: +55 (27) 999312189 - e-mail: renatpra@gmail.com.**

RESUMO

O estudo do método SVM mostra-se uma ferramenta que pode auxiliar na otimização operacional das empresas prestadoras de serviços. As perdas físicas na distribuição de água provocada por vazamentos devem ser rapidamente identificadas e evitadas. Os consumidores, agências reguladoras e poderes concedentes estão cada vez mais exigentes e buscam atendimentos rápidos e eficientes, provocando o desenvolvimento de novos métodos mais ágeis para garantir a efetividade dos atendimentos. O abastecimento de água, essencial à vida, é um serviço prestado por companhias de saneamento públicas e privadas, e os vazamentos, além de dispendiosos, provocam desabastecimentos e devem ser diminuídos para melhorar a prestação do serviço. O objetivo desse trabalho é apresentar um estudo com aplicação do método SVM (*Suport Vector Machine*) para reconhecimento de vazamentos a partir de variáveis operacionais em um setor de abastecimento de Viana, no estado do Espírito Santo, podendo sua aplicação se desdobrar a qualquer setor de abastecimento e também outros tipos de falhas. O estudo foi desenvolvido com amostras de dados do ano 2019, da Companhia Espírito Santense de Saneamento (CESAN) com Médias, Desvio Padrão e Intervalo Interquartil dos dados de vazão, nível de reservatório e pressão em um ponto de monitoramento. Foi identificado um valor satisfatório para avaliar dados operacionais que apresentam maior probabilidade de vazamentos no setor em comparação com dias sem vazamentos, permitindo assim a rápida identificação de número elevado de vazamentos com informações operacionais controladas por telemetria.

PALAVRAS-CHAVE: Vazamentos, SVM, Rede de Distribuição de Água

INTRODUÇÃO

As redes de distribuição de água (RDAs) constituem a maior parte do investimento inicial das companhias de saneamento. Além das tubulações, outros dispositivos, como válvulas, bombas e reservatórios são fundamentais para a boa operação do sistema, atendendo aos consumidores em quantidade e qualidade, ou seja, dentro de limites normativos de pressão e velocidade de escoamento (SWAMME E SHARMA, 2008).

Nos últimos anos as companhias de saneamento têm investido largamente na automação de seus sistemas de abastecimento de água, proporcionando dados em tempo real de vazão e pressão das cidades visando melhor otimização dos sistemas. Várias são as vantagens da obtenção de um modelo de identificação de falhas operacionais, entre as quais podem ser citadas a identificação de falas com modelos confiáveis e estimativa dos "*set points*" de controle de pressão e vazão ótimos ao longo do dia;

Estudos sobre comportamento do consumo de água em cidades comprovam sua característica dinâmica, multivariável e não linear, com fatores que influenciam de forma diferente em cada país, cidade e zonas de distribuição (FALKENBERG, 2016).

Diante do grande número de variáveis e falta de linearidade entre elas, torna-se necessário um método que simplifique e identifique de forma ágil a possibilidade de vazamentos que podem causar desabastecimento e perdas no sistema, bem como crie padrões operacionais para garantir a continuidade do abastecimento (SILVA JUNIOR, 2017).

Segundo Hillier e Lieberman (2010) a pesquisa operacional está relacionada à pesquisa sobre atividades de conduzir e coordenar operações. Ao identificar a existência de um problema, inicia-se o desenvolvimento do método científico através de levantamento de dados relevantes para a construção do modelo matemático, que irá alinhar os dados reais para obter a solução ótima.

Na pesquisa operacional, diversas técnicas de modelagem auxiliam no processo decisório para a identificação de falhas e o método SVM é uma delas. O algoritmo SVM, abreviação de *Suport Vector Machine* (Máquinas de Vetores de Suporte), é um algoritmo supervisionado de aprendizado de máquina aplicável para problemas de classificação. Criado por (VAPNIK, 1995), é um método de aprendizagem supervisionado usado para estimar uma função que classifique dados de entrada em duas classes. O objetivo do treinamento através de SVMs é a obtenção de hiperplanos que dividam as amostras de tal maneira que sejam otimizados os limites de generalização. É uma técnica útil para classificação de dados e seus fundamentos foram desenvolvidas por Vapnik (1998) . O SVM pertence a uma classe de algoritmos de aprendizado de máquina que são baseados em classificadores lineares com funções denominadas " Kernels". O objetivo da classificação de vetores de suporte é conceber uma maneira computacionalmente eficiente de aprender a separação de hiperplanos em um espaço de recursos de alta dimensão, onde 'Bons' hiperplanos são aqueles que otimizam os limites de generalização (CRISTIANINI,2000).

OBJETIVO

O presente trabalho tem como foco o estudo de identificação de vazamentos em um sistema de distribuição com população aproximada de 6500 habitantes, 2348 ligações e vazão média diária de 43 l/s em um setor predominantemente urbano. Variáveis operacionais serão utilizadas como preditores para identificação de vazamentos críticos com utilização de inteligência artificial. Logo, o objetivo deste trabalho é identificar quantidade de vazamentos significativos a partir das variáveis operacionais de uma determinada zona de distribuição com o método SVM.

JUSTIFICATIVA

Segundo SNIS (Sistema Nacional de Informações sobre o Saneamento), o consumo médio de água no país, em 2019, foi de 153,9 L/hab.dia apresentando um aumento de 0,6% em comparação a 2018. Por sua vez, ao distribuir água para garantir tal consumo, os sistemas sofreram perdas na distribuição, que na média nacional alcançaram 39,2%, 0,7 pontos percentuais acima do calculado em 2018. Possíveis causas para tal comportamento podem ter origem tanto na qualidade dos dados informados para o cálculo do indicador, quanto no efetivo aumento do volume de perdas por alguma ineficiência por parte dos prestadores de serviços. Dados recentes das principais companhias nacionais são apresentados na Figura 1. O Espírito Santo encontra-se com porcentagem próxima à média nacional (ES = 37,3%, BR = 39,2%).

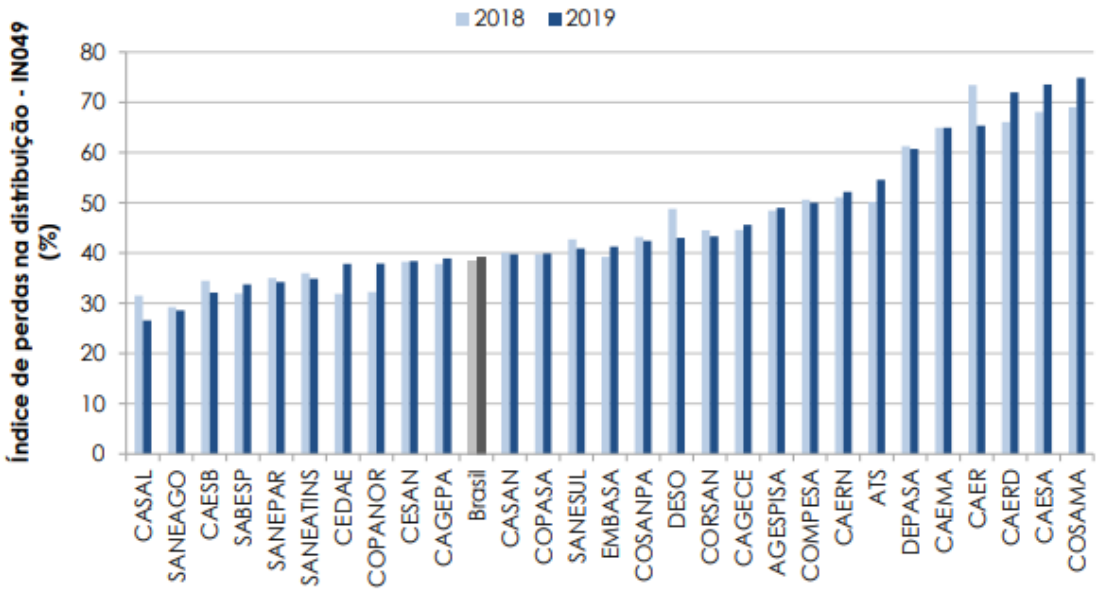


Figura 1: Gráfico de Perdas no Brasil. Ano 2019.

As perdas reais referem-se a toda água disponibilizada para distribuição que não chega aos consumidores. Essas perdas acontecem por vazamentos em adutoras, redes, ramais, conexões, reservatórios, falhas nos sistemas de medição e ligações clandestinas. O excesso de pressão, habitualmente em locais com grande variação topográfica e estado de conservação das redes são associados às perdas físicas, principalmente os vazamentos. A Tabela 1 apresenta as perdas reais, causas e efeitos (SNIS, 2019).

Tabela 1: Causa e efeito de perdas reais

	Perdas Reais
Tipo de ocorrência mais comum	Vazamento
Custos associados ao volume de água perdido	Custo de produção
Efeitos no Meio Ambiente	Desperdício do Recurso Hídrico
	Necessidades de ampliações de mananciais
Efeitos na Saúde Pública	Risco de contaminação
Empresarial	Perda do Produto
Consumidor	Imagem negativa (ineficiência e desperdício)
Efeitos no Consumidor	Repasse para tarifa
	Desincentivo ao uso racional

METODOLOGIA

Os dados do artigo foram fornecidos pela Companhia de abastecimento estadual com leituras das vazões (L/s), pressões (mca), e níveis de reservatórios (%) foram organizadas por dia e hora, relativos ao ano 2019. Havia cerca de três leituras, a cada hora , adquiridas por sensores e transmitidas remotamente para central de dados armazenadas em supervisório da companhia. Foi possível analisar histogramas dos dados organizados por dia, hora, semana, mês, e assim por diante.

Os dados vazão de distribuição, nível de reservatório e pressão de sucção em uma elevatória foram combinados com a intenção de definir o conjunto de variáveis de entradas para o classificador que melhor identificasse o perfil operacional que represente vazamentos na rede de distribuição. A quantidade de reclamações vazamentos no dia foi

utilizada como variável de saída do classificador. A partir de consulta do histórico de reclamações no SICAT (Sistema Integrado de Comercialização e Atendimentos da CESAN), foi possível obter as informações de vazamentos e suas características como dia e tipo de ocorrência.

As matrizes de atributos para simulações foram construídas com os dados operacionais de média diária e seus respectivos desvios padrão e Intervalos interquartil conforme apresentado na Tabela 2 relativa aos dados operacionais captados por sensores e transmitidos remotamente por telemetria via rádio frequência para a central de armazenamento de dados da Companhia.

Tabela 2: Parâmetros e Técnicas Analíticas Utilizadas.

Ponto de Medição	Dado	UNIDADE
Reservatório	Nível	%
Macromedidor Centro	Vazão de distribuição Setor Viana	L/s
EEAT Bom Pastor	Pressão de sucção na EEAT	mca

Por meio da análise de histograma, conforme Figura 2, foram considerados como críticos os dias com 3, 4, 5 ou mais relatos de vazamentos e suas recorrências no ano representados pela soma de vazamentos em ramais, redes e cavaletes relatados no dia..

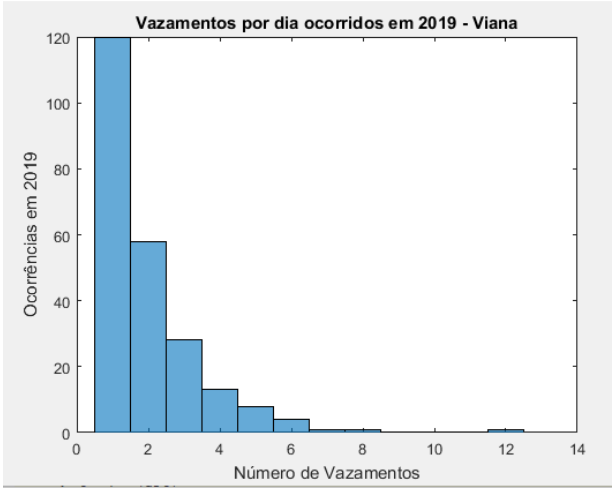


Figura 1: Histograma de Vazamentos por dia ocorridos em 2019 no setor

Para melhor entendimento da relação entre os atributos e sua influência na representação dos dados, foi aplicada a análise de Componentes Principais (ACP) ou *Principal Component Analysis* (PCA). O Método PCA é normalmente utilizado para reduzir as dimensões da matriz atributos. É um procedimento que utiliza uma transformação ortogonal (ortogonalização de vetores) para converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas chamadas de componentes principais.(FACELI et al., 2011)

A matriz atributos, foi reduzida a duas componentes principais, a partir da função PCA do MATLAB conforme equação 1. Sendo 'X 'a matriz atributos,

[COEFF, SCORE] = pca(X) equação (1)

Sendo assim, a partir dos dados relativos ao ano 2019, foram escolhidas aleatoriamente amostras de dias críticos e amostras de dias não críticos para realizar o teste de classificação. Os dias não críticos foram aqueles que não tiveram reclamações de vazamentos no dia.

Em cada caso, houve redução da matriz dados para duas componentes principais, ou seja, um vetor com SCORE 1 e SCORE 2. A classe predita, número de vazamentos, foi representada por um vetor sendo as linhas de classificação 1, dia crítico e 0 dia não crítico.

O classificador SVM foi utilizado como método de verificação. Uma máquina de vetores de suporte é um algoritmo que para um conjunto de métodos de aprendizado supervisionado analisam os dados e reconhecem padrões, usado

para classificação e análise de regressão. O SVM padrão toma como entrada um conjunto de dados e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico. Tem como objetivo encontrar uma função que produza saídas contínuas para os dados de treinamento que desviem no máximo de ε de seu rótulo desejado. Essa função deve ser o mais uniforme e regular possível (FACELI, 2011). As equações 2 e 3 apresentam funções para aproximar os pares (x_i , y_i) com uma precisão de ε . O vetor w é normal ao hiperplano cuja função tende a maximizar as distâncias dos pontos de diferentes classes conforme Figura 2.

$$\underset{\mathbf{w},b,\boldsymbol{\xi},\bar{\boldsymbol{\xi}}}{\text{Minimizar}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{n}\xi_i + \bar{\xi}_i\right)$$

equação (2)

Com as restrições:

$$\begin{cases} y_i - \mathbf{w}\cdot\mathbf{x}_i - b \leqslant \varepsilon + \xi \\ \mathbf{w}\cdot\mathbf{x}_i + b - y_i \leqslant \varepsilon + \bar{\xi} \\ \xi_i, \bar{\xi}_i \geqslant 0 \end{cases}$$

equação (3)

Nas equações, apresentadas acima, ξ_i representam as variáveis de folga e C uma constante que permitem lidar com ruídos e *outliers* nos objetos (FACELI, 2011). .

Os dados do conjunto de exemplos de treinamento foram marcados em duas categorias (dias críticos e não críticos). O SVM construiu um modelo que atribui novos exemplos a uma categoria ou outra. Estes exemplos foram então mapeados no mesmo espaço e preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles foram colocados.

Em outras palavras, o que o SVM faz é encontrar uma linha de separação, mais comumente chamada de hiperplano entre dados das duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes, conforme figura 2:

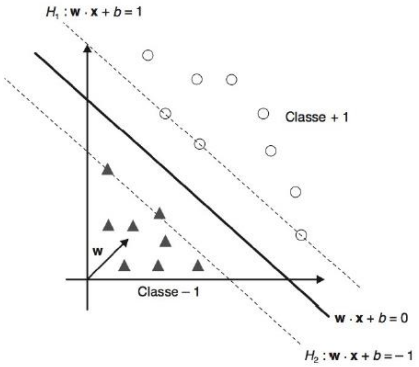


Figura 2: Ilustração de hiperplanos canônicos e separador. Fonte (FACELI,2011).

Neste trabalho, o teste foi realizado por meio de validação cruzada com 5 conjuntos, onde 4 serviram como treinamento e 1 como teste. A média dos 5 testes ofereceu a acurácia do classificador. A análise final do melhor teste foi realizada por matriz confusão e curva ROC. A Curva Característica de Operação do Receptor (Curva COR), ou, do inglês, *Receiver Operating Characteristic Curve* (ROC *curve*), é uma representação gráfica que ilustra o desempenho de um sistema classificador. A análise da Curva ROC, equivale à área abaixo da curva (*Area under the curve*): Ela varia de 0 a 1, sendo que 0,5 representa um modelo que seria completamente aleatório. Quanto mais próximo de 1, maior a probabilidade de classificação correta.

RESULTADOS

Para identificação da quantidade de vazamentos críticos, onde os dias com mais ocorrências considerados críticos, foram propostos cenários com a intenção de investigar os melhores resultados do classificador para as amostras aleatórias. Os testes contaram com atributos de média diária e desvio padrão diário e intervalo interquartil dos dados, porém com redução de dimensão por PCA dos dados de nível do reservatório, vazão de distribuição e pressão de sucção da Elevatória.

Os resultados estão organizados nas tabelas 3, 4 e 5, para classificação de mais de três, quatro e cinco vazamentos em comparação aos dias sem relatos de vazamento.

Tabela 3: Resultados do classificador para mais de três vazamentos em comparação com dias sem vazamentos

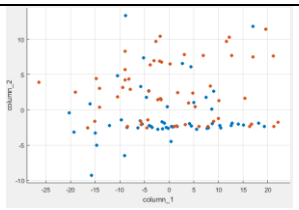
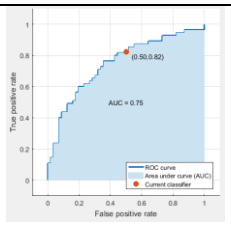
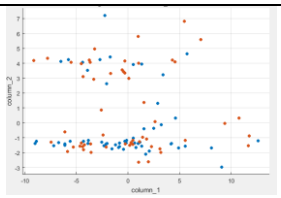
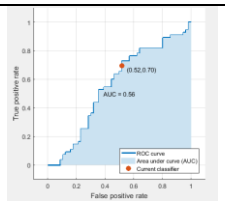
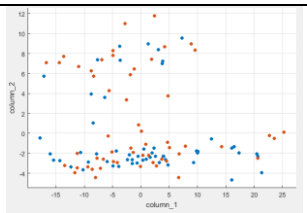
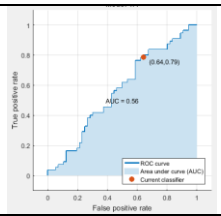
Acima de 3 vazamentos 56 amostras	Scatterplot	Curva ROC	Resultados
Atributo preditor Média diária			69,6% de acurácia 0,75 AUC (área abaixo da curva ROC)
Atributo preditor Desvio Padrão diário			58,9% de acurácia 0,56 AUC (área abaixo da curva ROC)
Atributo preditor Intervalo Interquartil diário			57,1% de acurácia 0,56 AUC (área abaixo da curva ROC)

Tabela 4: Resultados do classificador para mais de quatro vazamentos em comparação com dias sem vazamentos

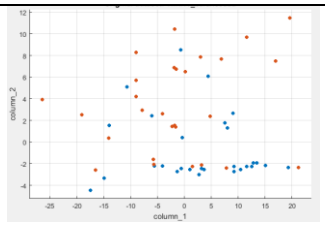
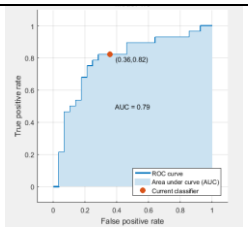
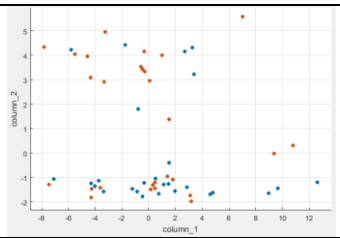
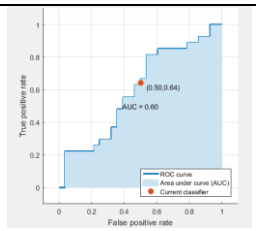
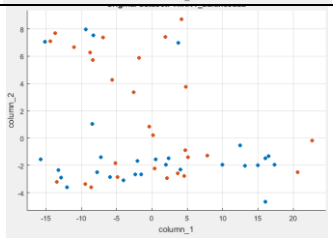
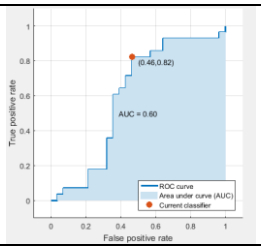
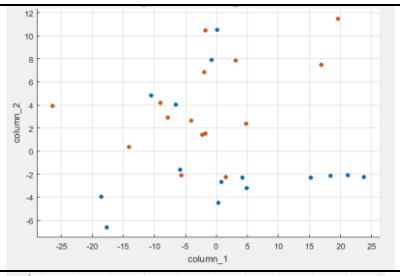
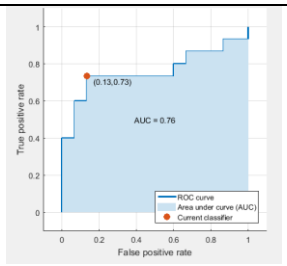
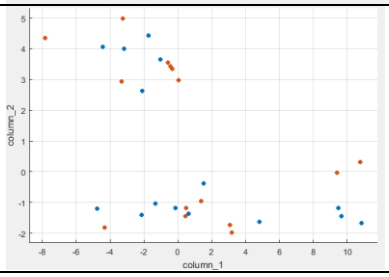
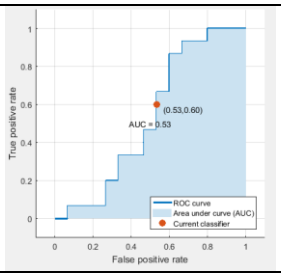
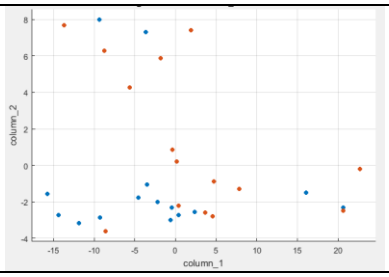
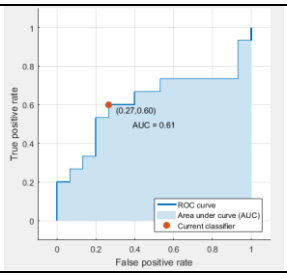
Acima de 4 vazamentos 28 amostras	Scatterplot	Curva ROC	Resultados
Atributo preditor Média diária			73,2% de acurácia 0,79 AUC (área abaixo da curva ROC)
Atributo preditor Desvio Padrão diário			58,9% de acurácia 0,56 AUC (área abaixo da curva ROC)
Atributo preditor Intervalo Interquartil diário			57,1% de acurácia 0,56 AUC (área abaixo da curva ROC)

Tabela 5: Resultados do classificador para mais de cinco vazamentos em comparação com dias sem vazamentos

Acima de 5 vazamentos 15 amostras	Scatterplot	Curva ROC	Resultados
Atributo preditor Média diária			80,0% de acurácia 0,76 AUC (área abaixo da curva ROC)
Atributo preditor Desvio Padrão diário			53,3% de acurácia 0,53 AUC (área abaixo da curva ROC)

Atributo preditor Intervalo Interquartil diário			66,7% de acurácia 0,61 AUC (área abaixo da curva ROC)

CONCLUSÕES

A técnica de PCA auxiliou a análise dos resultados já que, para investigação multivariada, com a redução do número de variáveis originais do problema foi útil para identificação visual através de gráficos tipo Scatter Plot dos melhores cenários. Os resultados da área abaixo da Curva ROC, com o classificador SVM, apresentaram melhor sensibilidade quando aplicados os valores preditores de média diária dos dados operacionais..

As classes das ocorrências de vazamentos foram 0 e 1, sendo, 0 os dias sem relatos de vazamentos. A classe 1, representa os dias com vazamentos e criados três cenários diferentes. Em 2019, foram 56 dias com mais de 3 vazamentos, 28 dias com mais de quatro vazamentos e 15 dias com 5 ou mais vazamentos, dias considerados críticos no setor estudado.

O melhor resultado encontrado neste trabalho para o sistema em questão foi a predição de dias com mais de 4 vazamentos , que aplicou a média diária dos atributos. A partir deste indicador, o operador poderá identificar padrões operacionais de tal forma a evitar vazamentos, identificar desvios e corrigir eventuais falhas.

Os resultados podem ser considerados satisfatórios, levando-se em conta que a base de dados utilizada para treinamento é pequena. Nos próximos trabalhos, poderá ser feita a avaliação pontual dos maiores desvios para aprimorar a classificação e identificar os tipos de falhas.

Com relação à identificação de falhas operacionais, visto que as cidades apresentam um comportamento dinâmico ao longo dos anos, o modelo deve ser treinado periodicamente para adaptar-se às mudanças do meio buscando melhoria contínua, afim de controlar a distribuição com o mínimo de vazamentos.

Como sugestões a trabalhos futuros, recomenda-se identificar pontualmente os vazamentos suas características a partir da matriz atributos as e utilizar diferentes classificadores como Redes Neurais Artificiais, além de realizar o mesmo teste para outros sistemas com diferentes características.

REFERÊNCIAS BIBLIOGRÁFICAS

1. CUNHA, Maria da Conceição; SOUSA, Joaquim. Hydraulic infras tructures design using simulated annealing. Journal of Infrastructure Systems, v. 7, n. 1, p. 32-39, 2001.
2. CRISTIANINI AND SWHAWE-T . (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
3. FACELI, K., Ana, L., GAMA, J., & CARVALHO, A. (2011). Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina (LTC (ed.); 1o Edição).
4. FALKENBERG, A. V., DYMINSKI, A. S., & RIBEIRO, E. P. (2016). Redes Neurais Artificiais Aplicadas à Previsão de Consumo de Água. 319–324. https://doi.org/10.21528/cbrn2003-068
5. HILLIER, F. S.; LIEBERMAN, G. J. Introdução à pesquisa operacional. 9. ed. Porto Alegre: AMGH, 2013.
6. SILVA JUNIOR, J. F. (2017). Detecção de Perdas em Sistemas de Distribuição de Água através de Redes de Sensores Sem Fio. Universidade Federal de Pernambuco.
7. SWAMEE, Prabhata K.; SHARMA, Ashok K. Design of water supply pipe networks. John Wiley & Sons, 2008.
8. VAPNIK, V. (1998) Statistical Learning Theory. John Wiley & Sons, Chichester.