UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS À PREDIÇÃO DE VAZAMENTOS EM RAMAIS DE REDES DE DISTRIBUIÇÃO DE ÁGUA

CRISTIANO GONÇALVES NASCIMENTO GOUVEIA

ORIENTADOR: ALEXANDRE KEPLER SOARES

SEMINÁRIO DE DISSERTAÇÃO DE MESTRADO EM TECNOLOGIA AMBIENTAL E RECURSOS HÍDRICOS

BRASÍLIA/DF: FEVEREIRO – 2022
UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS À PREDIÇÃO DE VAZAMENTOS EM RAMAIS DE REDES DE DISTRIBUIÇÃO DE ÁGUA

CRISTIANO GONÇALVES NASCIMENTO GOUVEIA

DISSERTAÇÃO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM TECNOLOGIA AMBIENTAL E RECURSOS HÍDRICOS.

APROVADA POR:

_____

Prof. Alexandre Kepler Soares, PhD. (ENC-UnB)
(Orientador)

_____

Prof. Carlos Henrique Ribeiro Lima, PhD. (ENC-UnB)
(Examinador Interno)

_____

Prof. Bruno Melo Brentan, PhD. (EHR-UFMG)
(Examinador Externo)

BRASÍLIA/DF, 11 DE FEVEREIRO DE 2022

FICHA CATALOGRÁFICA

GOUVEIA, CRISTIANO GONÇALVES NASCIMENTO
Técnicas da Aprendizado de Máquina aplicadas à predição de vazamentos em ramais de redes de distribuição de água [Distrito Federal] 2021.
Xvii, 116p., 210 x 297 mm (ENC/FT/UnB, Mestre, Tecnologia Ambiental e Recursos Hídricos, 2022).
Dissertação de Mestrado – Universidade de Brasília. Faculdade de Tecnologia.
Departamento de Engenharia Civil e Ambiental.
1.Sistemas de abastecimento de água 2.Gestão da infraestrutura
3.Perdas de água 4.Técnicas de Aprendizado de Máquina
I. ENC/FT/UnB II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

GOUVEIA, C. G. N. (2022). 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, 116p.

CESSÃO DE DIREITOS

_____

Cristiano Gonçalves Nascimento Gouveia

Gouveia.crs@gmail.com.

À minha esposa e meu filho, por todo o amor.

"*Are we heading towards ecological disaster or technological paradise? (...)*

In a few decades, people will look back and think the answer to all these questions were obvious."

Yuval Noah Harari, *Sapiens: A Brief History of Humankind*.

RESUMO

TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS À PREDIÇÃO DE VAZAMENTOS EM RAMAIS DE REDES DE DISTRIBUIÇÃO DE ÁGUA

As perdas de água em sistemas de abastecimento são um desafio aos prestadores de serviços de saneamento dado que as ações para gestão da infraestrutura e perdas demandam recursos (humanos, financeiros, tecnológicos e materiais) significativos para que investimentos eficientes e assertivos sejam aplicados. No que tange as perdas reais de água, um conceito amplamente aplicado é o Controle Ativo de Vazamentos, que possui em uma de suas componentes a pesquisa de vazamentos não visíveis nas redes de distribuição e em ramais de ligação à unidade consumidora. A pesquisa de vazamentos habitualmente é aplicada extensivamente na infraestrutura, dado que é necessário realizar inspeções para a identificação de vazamentos. Esta condição é dispendiosa por natureza. Portanto, meios para direcionamento das investigações sobre perdas reais têm sido amplamente estudados para direcionar e envidar esforços com maior produtividade e eficiência. No âmbito das pesquisas para proposição de ferramentas nesta temática, técnicas de Aprendizado de Máquina compõem métodos que se mostram interessantes para auxiliar a compreensão do tomador de decisão quanto aos fatores intervenientes ao surgimento de vazamentos, o que tem corroborado, conforme revisão bibliográfica realizada, com o aprimoramento dos processos de gestão da infraestrutura e controle de perdas. Neste contexto, a presente pesquisa aplicou à uma base de dados da Caesb, consistida por 38 variáveis (operacionais, físicas, ambientais e ambientais), 12 modelos de classificação por Aprendizado de Máquina (*Linear SVM*, *Radial SVM*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, *Random Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*), processados em Python 3.7 e bibliotecas para Aprendizado de Máquina e Ciência de Dados. Os resultados obtidos identificam que modelos do tipo *Ensemble Learning* performaram melhor, com mais destaque para o *AdaBoost*, obtendo acurácia final de 59,70% para toda a base de dados após hiper-parametrização. O processamento da base de dados incluindo variáveis hidráulicas obtidas por meio de simulação das redes de distribuição incrementou em média 2,03% de acurácia, indicando que tais componentes agregam valor à análise preditiva de vazamentos. A discretização espacial dos dados por área de atendimento de reservatório apoiado permitiu a obtenção de melhores acurácias, obtendo-se acurácias de até 63,6% em algumas regiões. A

avaliação da significância das variáveis preditoras (pressões operacionais, material das tubulações, tipo de solo sob os tubos, idade da rede e do ramal de serviço, declividade do solo, e outras variáveis) permite a avaliação da dinâmica da falha (vazamento), fornecendo informações sobre as condições de maior vulnerabilidade, podendo-se priorizar ações para Controle Ativo de Vazamentos (priorização da infraestrutura para ações de Pesquisa de Vazamentos) e Gestão de Ativos/Reabilitação de Infraestrutura (priorização de tubulações para substituição ou reabilitação).

ABSTRACT

MACHINE LEARNING TECHNIQUES APPLIED TO LEAKAGE PREDICTION IN WATER DISTRIBUTION NETWORK SERVICE CONNECTIONS

Non-Revenue Water in water supply systems is a challenge for sanitation service providers. It requires several actions and resources (human, financial, technological, and material) for efficient and assertive investments to cope with infrastructure deterioration and water leakage. Regarding physical water losses, a widely applied concept is the Active Leakage Control, which demands actions to find non-visible leaks in the distribution networks and water service connections. Leak localization is usually extensively applied to infrastructure to find them, but this process is time-consuming and field inspections are costly. Therefore, means to drive investigations on physical losses have been studied to direct and carry out efforts to enhance productivity and efficiency to find and repair leaks. In the context of research and tool propositions on this agenda, Machine Learning techniques are promising to support the decision maker's comprehension of predictive factors to water leakage. According to the literature review carried out, Machine Learning models can improve infrastructure asset management and water loss control processes. In this setting, this research applied to Caesb's infrastructure database, consisting of 38 variables (operational, physical, environmental, and environmental), 12 classification models by Machine Learning (Linear SVM, Radial SVM, Logistic Regression, KNN, Decision Tree, Naive Bayes, Random Forest, bagged KNN, bagged Decision Tree, AdaBoost, Gradient Boosting, e XGBoost). Data processing was done through Python 3.7 and libraries for Machine Learning and Data Science. Ensemble Learning models performed better, and AdaBoost obtained 59,70% as final score after hyper-parametrization for the entire database. Hydraulic variables contributed to increase an average of 2.03% of accuracy, indicating that such components add value to the predictive analysis of leaks. Data organized by distribution network area obtained accuracies up to 63.6%. The evaluation of predictor variables (operating pressures, material of the pipes, type of soil under the pipes, age of the network, soil slope, and other variables) can provide information on the most vulnerable conditions, giving priority to actions for Active Leakage Control and Asset Management/Infrastructure Rehabilitation.

SUMÁRIO

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE SÍMBOLOS, NOMENCLATURAS E ABREVIAÇÕES

ACP	Análise dos Componentes Principais
AESBE	Associação Brasileira das Empresas Estaduais de Saneamento
AWWA	*American Water Works Association*
CAESB	Companhia de Saneamento Ambiental do Distrito Federal
COR	Característica de Operação do Receptor
DMC	Distrito de Medição e Controle
ELL	*Economic Level of Leakage*
EPAL	Empresa Portuguesa das Águas Livres
FP	Falso Positivo
FN	Falso Negativo
IWA	*International Water Association*
KNN	*K-nearest Neighbor*
NBR	Norma Brasileira
PCA	*Principal Component Analysis*
RAP	Reservatório Apoiado
ROC	*Receiver Operating Characteristic Curve*
SIG	Sistema de Informações Geográficas
SNIS	Sistema Nacional de Informações Sobre Saneamento
SVM	*Support Vector Machine*
SVC	*Support Vector Classifier*
SVR	*Support Vector Rregressor*
TFP	Taxa Falso Positivo
TFN	Taxa Falso Negativo
TVP	Taxa Verdadeiro Positivo
TVN	Taxa Verdadeiro Negativo
t-SNE	*t-distributed stochastic neighbor embedding*
t-SVD	*Truncated Singular Value Decomposition*
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
VRP	Válvula Redutora de Pressão

1. INTRODUÇÃO

As Redes de Distribuição de Água são um elemento fundamental nos sistemas urbanos de abastecimento de água, sendo o componente mais oneroso, constituindo a parte da infraestrutura que permite o acesso ao serviço prestado com qualidade e quantidade à porta de cada cliente. Como característica natural da infraestrutura, há uma deterioração progressiva das condições físicas e de desempenho ao longo dos anos; inclusive daquelas projetadas, construídas e operadas adequadamente (Rajani e Kleiner, 2002). Consequentemente, como lidar com a deterioração dos sistemas de distribuição de água é uma questão importante à medida que a infraestrutura envelhece, chegando gradativamente ao fim de sua vida útil. Portanto, são necessários investimentos em manutenção e reabilitação dos elementos constituintes das redes de distribuição de água (Giustolisi *et al.*, 2006), buscando-se a sustentação de níveis adequados da prestação do serviço, considerando aspectos econômico-financeiros, ambientais, operacionais, regulatórios e sociais.

A deterioração progressiva da infraestrutura e seu declínio em termos de performance podem ser apropriadas por meio da avaliação das perdas de água em um sistema (Farley e Trow, 2003). As perdas de água são divididas em perdas reais e perdas aparentes (Hirner e Lambert, 2000; Alegre *et al*., 2017). Segundo Hirner e Lambert (2000), as perdas reais ocorrem em tubulações, reservatórios, e em ramais de ligação às instalações dos clientes. A perda real não é aproveitada nem pela população, nem pela concessionária em termos de uso especial. As perdas aparentes constituem perdas comerciais e são relacionadas à submedição em hidrômetros, falhas na gestão de dados comerciais ou emissão de contas, fraudes e ligações clandestinas (uso não autorizado). Com impacto direto, as perdas provocam desperdício de água, instabilidade técnica dos componentes da rede, deterioração da qualidade da água, iniquidade ao acesso ao serviço, aumento dos custos de operação e manutenção, além da perda de receitas necessárias para sustentar e expandir o acesso à água tratada e de qualidade. Em termos de impactos indiretos, as perdas reais podem corroborar com o agravamento de conflitos de uso de recursos hídricos em bacias hidrográficas que percebem tal situação, por exemplo.

O envelhecimento dos elementos da rede de distribuição potencializa o surgimento de rompimentos e vazamentos em tubulações e em ramais, induzindo ao aumento das perdas

reais. Tais vazamentos podem ser classificados em vazamentos visíveis e não visíveis (Alegre *et al*., 2017). Os vazamentos visíveis consistem em situações em que a vazão do vazamento aflora e pode ser facilmente detectável. (Hirner e Lambert, 2000). Os não visíveis são eventos em que a vazão é muitas vezes baixa e não alcança a superfície do terreno, sendo que a sua detecção demanda investigações específicas para localização (Lambert *et al*., 1999; Farley *et al*., 2008; Mutikanga *et al*., 2013). Embora um vazamento visível, em termos gerais, tenha uma vazão instantânea elevada, são rapidamente corrigidos. Entretanto, vazamentos não visíveis podem vazar durante dias, meses ou até anos caso não sejam identificados e corrigidos (Farner e Thornton, 2005). Por isso, devido ao longo tempo de reparo, os volumes perdidos neste segundo caso são superiores.

Assim, meios de analisar a infraestrutura, diagnosticar as perdas de água, avaliar alternativas e orientar o processo decisório para a aplicação de recursos com vistas à eficiência dos sistemas de abastecimento de água é pauta amplamente estudada em termos de indicadores, manuais e procedimentos (Alegre *et al*., 2000; AWWA, 2009; Lambert *et al*., 2014; *European Comission*, 2015; AESBE, 2015; AWWA, 2016). Em avanço a essa pauta, em termos de perdas reais, nos últimos anos vários estudos versam sobre como melhorar a compreensão e a predição sobre os fatores intervenientes ao processo de quebra e vazamento em redes de distribuição de água, não apenas por meio de métodos e modelos determinísticos e probabilísticos, mas também por meio de modelos baseados em aprendizado de máquina para classificação ou regressão de dados (Snider e McBean, 2020). Tais abordagens preditivas podem impulsionar os métodos de gestão das perdas reais em sistemas de abastecimento de água, dado que diversos estudos têm apontado resultados e acurácia superiores por meio da aplicação de modelos de Aprendizado de Máquina em detrimento aos modelos estatísticos aplicados à predição de rompimento em redes de abastecimento de água (Robles-Velasco *et al*., 2020; Giraldo-González e Rodrígues, 2020; Snider e McBean, 2020b).

Em uma companhia de saneamento, os recursos para investimento em reabilitação da infraestrutura e gestão das perdas de água competem com outras pautas, tais como expansão da cobertura do serviço. Portanto, à medida que ferramentas estejam disponíveis para subsidiar o investimento prudente e assertivo, deve-se aplicá-las com vistas a majorar a eficiência dos serviços prestados. Neste sentido, 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-se a possibilidade de direcionamento de ações para o controle de perdas de água por meio da aplicação de modelos de Aprendizado de Máquina. A Revisão Bibliográfica desta pesquisa identificou 28 publicações internacionais sobre este tema, focadas em quebra de tubos da rede de distribuição propriamente dita, observando-se a oportunidade de explorar tal abordagem, também, aplicada aos ramais de ligação às unidades consumidoras da água tratada ofertada pelos prestadores de serviço.

Estão à disposição uma série de métodos de Aprendizado de Máquina que podem ser comparados em termos de desempenho, buscando-se identificar e aplicar aqueles com melhor performance para contribuição ao processo decisório. Assim, a presente pesquisa aplicou e comparou o desempenho prestado pelos seguintes modelos: *Linear SVM*, *Radial SVM*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, *Random Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*. Portanto, o escopo da presente pesquisa também contempla a avaliação da performance dos modelos, além da identificação das variáveis preditoras à falha, e a avaliação da aplicabilidade dos resultados obtidos nos processos de gestão de perdas de água.

A aplicação de tais métodos foi processada utilizando-se bases de dados da Companhia de Saneamento Ambiental do Distrito Federal (Caesb), integrando-se dados cadastrais e operacionais, com suporte de Sistemas de Informação Geográfica (SIG) e modelos hidráulicos de redes de distribuição. Os dados utilizados para predizer vazamento em ramais das redes de distribuição foram agrupados nas seguintes classes: aspectos operacionais, aspectos físicos, aspectos comerciais, e aspectos ambientais, perfazendo 38 variáveis. Segundo Alizadeh *et al*. (2019), Robles-Velasco *et al*. (2020), e Snider & McBean (2020b), as variáveis obtidas por meio da dinâmica operacional das redes de distribuição são pertinentes quanto à falha por rupturas e vazamentos. Assim do total de variáveis, seis foram obtidas por meio de simulação hidráulica das redes de distribuição, considerando que as condições operacionais em termos de pressões, velocidades e perda de carga podem influenciar o processo de deterioração e falha dos tubos, constituindo variáveis relevantes à predição.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Avaliar o potencial de aplicação de modelos de Classificação por Aprendizado de Máquina para auxílio aos processos de Gestão de Ativos e Controle de Perdas em sistemas de abastecimento de água, utilizando-se dados sobre vazamentos e condições de contorno da infraestrutura (variáveis operacionais, físicas e ambientais), aplicando-se como estudo de caso redes da Companhia de Saneamento Ambiental do Distrito Federal-Caesb.

2.2. OBJETIVOS ESPECÍFICOS

Avaliar a performance de diferentes modelos de Aprendizado de Máquina para predizer vazamentos em ramais da rede de água;

Avaliar a relevância dos fatores contribuem com a falha nos ramais das redes de distribuição de água estudadas;

Avaliar a possibilidade de incremento na performance dos modelos de Aprendizado de Máquina quando incluídas as variáveis obtidas por meio de modelos hidráulicos;

Avaliar a possibilidade de melhoria da performance por meio de *Ensemble Learning Models* e por hiper parametrização; e,

Avaliar os resultados obtidos e possibilidade de aplicação nos processos de gestão das perdas de água.

3. REVISÃO BIBLIOGRÁFICA E FUNDAMENTAÇÃO TEÓRICA

3.1. DETERIORAÇÃO E FALHA EM REDES DE DISTRIBUIÇÃO DE ÁGUA

O envelhecimento dos tubos em redes de distribuição de água é iniciado assim que seu assentamento é realizado no subsolo, sendo a deterioração da infraestrutura concomitante ao estresse gerado pelas condições operacionais e ambientais impostas ao material do tubo. Quando a resistência à carga não suporta o estresse interno ou externo, ocorre a quebra do tubo. A deterioração estrutural é típica, mas não é o único tipo de deterioração do tubo. A deterioração pode ser classificada em duas categorias Rajani e Kleiner (2002): (01) deterioração estrutural, que diminui a capacidade de resistência estrutural residual do tubo, podendo-se culminar em quebra, rompimento ou ruptura; e, (02) deterioração não estrutural, caracterizada por maior rugosidade das superfícies internas dos tubos e diâmetro interno mais estreito, resultando em redução da capacidade hidráulica e degradação da qualidade da água.

A deterioração dos tubos em sistemas de abastecimento de água pode conduzir o sistema à falha. Neste contexto, uma falha pode ser definida de várias maneiras, incluindo: uma redução de pressão de serviço, abaixo de um mínimo especificado em condições operacionais normais (no caso brasileiro, a pressão dinâmica mínima a ser observada em redes de abastecimento de água é de 100 kPa, conforme NBR 12218:2017); uma interrupção não planejada do abastecimento; ou, um evento que leva a um impacto negativo na qualidade física, química ou biológica da água. Assim, antes de realizar uma análise de probabilidade de falhas em redes de distribuição, é importante conceituar a falha. Segundo Rajani e Kleiner (2002), a falha do tubo pode ser classificada em três grupos:

- Falha estrutural. A ruptura física possui simples definição: é a quebra (trinca) ou ruptura (separação, descontinuidade do tubo) originada na tubulação, demandando-se intervenção ativa para reparo. O foco das pesquisas sobre falha estrutural versa sobre a deterioração externa dos tubos. A falha estrutural do tubo desempenha um papel predominante sobre os demais aspectos de falha em redes de distribuição de água, pois a falha estrutural pode conduzir, também, às falhas hidráulicas e de qualidade da água.

- Falha hidráulica. Geralmente é definida como a incapacidade da rede de distribuição de água de atender à demanda de água com a pressão de serviço mínima especificada. Uma falha hidráulica pode ocorrer por vários motivos, incluindo: demanda do sistema superior ao previsto e projetado (consumo ou vazamentos); uma falha de componente físico (por exemplo, desligamento de um conjunto motobomba tipo *booster* de pressurização direta na rede); e deterioração severa da condição dos tubos, resultando em expressiva redução da capacidade hidráulica da rede.

- Falha da qualidade da água. As interações da água fornecida e do material do tubo podem conduzir a reações químicas e bioquímicas no sistema. A corrosão de tubo é um caso típico. Se as impurezas externas invadem o sistema por meio de algumas quebras, pode-se desencadear alterações nas características físicas, químicas ou biológicas da água. As falhas na qualidade são frequentemente classificadas com base na maneira pela qual a falha ocorreu: (01) entrada e intrusão de contaminantes por vazamentos e rachaduras no tubo; (02) crescimento de micro-organismos ao longo das paredes do tubo; (03) lixiviação de produtos químicos e produtos de corrosão nas paredes dos tubos; (04) permeação de produtos orgânicos a partir de componentes como juntas, na água; e, (05) consumo de todo o cloro residual acrescido à água tratada na estação de tratamento.

A falha estrutural apresenta duas formas típicas de falha operacional ao longo da estrutura física do tubo, variando de vazamento à ruptura. Os termos vazamento e ruptura são utilizados para diferenciar o nível de urgência e facilidade de diagnóstico entre as diferentes falhas dos tubos; assim, Mays (2000) caracteriza vazamento e ruptura conforme Tabela 3.1.

Tabela 3.1 - Caracterização de ruptura e vazamento em tubulações de água (Mays, 2000)

	Ruptura	Vazamento
Detecção	Facilmente detectável pelas condições do recobrimento do tubo e pressão da água	Difícil detecção. Normalmente depende de equipamentos específicos
Impacto no serviço	Alta probabilidade de provocar interrupção do serviço	Baixa probabilidade de causar interrupção do serviço
Ocorrência	Tipicamente ao longo do comprimento do tubo	Normalmente ocorre em acessórios das tubulações e em suas laterais
Urgência do reparo	Requer atenção imediata	Os reparos podem ser programados e não são urgentes

As falhas estruturais mais comuns em tubulações para abastecimento de água agrupadas por material, bem como proposta de classificação das falhas, são relacionadas nas Tabelas Tabela 3.2 e Tabela 3.3, respectivamente (InfraGuide, 2003).

Tabela 3.2 – Modo comum de falha estrutural em tubulações de SAA, segundo o material da tubulação (InfraGuide, 2003).

Material do tubo	Modos de falha estrutural
Ferro fundido	Rupturas longitudinais e transversais, rachadura da bolsa e corrosão por orifícios
Ferro fundido dúctil	Corrosão por orifícios
Aço	Corrosão por orifícios. Tubos de grande diâmetro são suscetíveis ao colapso
Policloreto de vinila	Rupturas longitudinais por estresse mecânico excessivo. Suscetível à ruptura devido frio extremo
Polietileno de alta densidade	Degradação mecânica por instalação incorreta, suscetíveis à vácuo e imperfeições comuns

Continuação da Tabela 3.2 – Modo comum de falha estrutural em tubulações de SAA, segundo o material da tubulação (InfraGuide, 2003).

Cimento amianto	Rupturas transversais, degradação por águas com características mais agressivas e rachaduras longitudinais
Concreto	Danos devido instalação imprópria, degradação da tubulação por solos agressivos.

Tabela 3.3 – Classificação de falhas estruturais de SAA (InfraGuide, 2003)

Grupo	Descrição
Eventos de causas naturais	Fogo, tempestade, inundação e terremoto. A previsão de tais eventos é desconhecida e incontrolável, mas sua probabilidade e gravidade podem ser estatisticamente previstos.
Impactos externos	resultado de ações de terceiros, como queda e energia, acidentes, greves. Esta fonte de risco é imprevisível; no entanto, as consequências podem ser mitigadas por planos de gestão.
Agressões externas	Atos deliberados de terceiros que resulta em destruição de bens. As consequências do fracasso podem ser reduzidas por meio de programas de segurança e de proteção às instalações estrategicamente importantes.
Envelhecimento da infraestrutura e deterioração física	A condição da infraestrutura e a sua degradação pode ser prevista e determinada. Esses fatores são categorizados em três grupos: os fatores físicos, fatores ambientais e fatores operacionais.
Risco de falha operacional	Esta categoria surge como resultado da forma como a infraestrutura é projetada, gerenciada e operada para atender aos objetivos organizacionais. Esse risco pode ser reduzido por meio da avaliação do estado e desempenho e inspeção de ativos em intervalos regulares e através de programas de manutenção preventiva.

3.1.1. Fatores intervenientes à falha

Existem vários fatores que podem impactar negativamente a condição de tubos enterrados, tais como as redes de distribuição de água. Kishawy e Gabbar (2010) resumiram os fatores que podem ameaçar a integridade das tubulações subterrâneas como (1) operação incorreta, (2) material da tubulação, (3) corrosão e rachaduras mecânicas, (4) forças terrestres, como terremotos e deslizamentos de terra, e (5) condições meteorológicas e fatores relacionados, como a temperatura. Assim, concluem que projetos para assentamento subterrâneo de tubulações não deve depender apenas de critérios de pressão e tensão, sendo imprescindível a compreensão e inclusão de considerações sobre as condições sob as quais as tubulações estarão expostas após implantação.

A deterioração das tubulações é resultado de fatores intervenientes de ordem operacional e, também, estáticos e dinâmicos. Fatores estáticos, como material, diâmetro e idade do tubo, são fixos ao longo do tempo; considerando que fatores dinâmicos, incluindo fatores ambientais e operacionais, como umidade do solo, temperatura, deslizamento de terra, tensões externas e pressões operacionais, podem mudar ao longo do tempo (Kleiner e Rajani 2001; Wang *et al*. 2009; Farmani *et al*. 2017).

Existem mecanismos físicos que podem levar à ruptura e falha da tubulação, como (1) cargas internas resultantes da pressão operacional e cargas externas devido ao tráfego e sobrecarga do solo; e (2) propriedades estruturais do tubo, interação tubo-solo e qualidade da instalação (Rajani e Kleiner, 2001). Além desses fatores, Makar *et al.* (2001) discorrem sobre a relevância das falhas de fabricação, forças excessivas e os erros humanos, que possuem impacto significativo no desempenho e na vida útil das redes de água. Em estudos estruturais e geotécnicos, Kamel e Meguid (2012) indicam que a perda de contato entre os dutos e o solo circundante pode aumentar a pressão de contato e, consequentemente, as tensões no material da tubulação. Kaddoura e Zayed (2018) examinaram o efeito de vazios de erosão em torno de tubos enterradas, avaliando-se como tais vazios podem reduzir a vida útil de uma tubulação. Eles identificaram cinco fatores que mais contribuem para a ocorrência de vazios em torno de tubulações: tipo de solo, tipo de leito, profundidade do duto, idade do duto e nível do lençol freático.

Agrupando-se os fatores intervenientes à falha em tubulações de sistemas de abastecimento de água, InfraGuide (2003) classificou os fatores que contribuem para a deterioração das tubulações de água em três categorias principais: físicos, ambientais e operacionais. De acordo com esta classificação, os fatores físicos incluem material do tubo, espessura da parede do tubo, idade do tubo, diâmetro do tubo, tipo de junta, revestimento do tubo, condições de instalação e fabricação, e outros. Tipo de solo, águas subterrâneas, clima, localização de tubo, correntes elétricas e atividades sísmicas são considerados como fatores ambientais, enquanto outros pesquisadores também incluíram chuva, tráfego e aterro da vala como fatores ambientais (Kabir *et al*. 2015).

A pressão interna da água, a pressão transiente, o vazamento, a qualidade da água, a velocidade do fluxo, e as práticas de operação e manutenção são exemplos de fatores operacionais (InfraGuide, 2003). Também há de se pontuar a natureza e data da última falha, informações do último reparo, qualidade da água e método de construção como fatores operacionais que afetam a taxa de falha das tubulações de água (InfraGuide, 2003). O Quadro 3.1 apresenta o agrupamento dos fatores em estáticos, dinâmicos e operacionais que motivam a falha de tubos em curto e longo prazo.

A avaliação dos fatores que afetam a condição de redes de abastecimento de água é essencial para que os prestadores de serviço possam: (1) desenvolver estratégias que mitiguem a probabilidade de falha da tubulação; (2) avaliar os custos resultantes da falha de uma tubulação, considerando a execução de reparos emergenciais, prejuízo à indicadores operacionais e regulatórios, sinistros a terceiros, interrupção do tráfego, entre outros; e, (3) planejar a substituição dos tubos, ponderando-se questões técnicas e econômicas.

O Quadro 3.1 – Fatores estáticos, dinâmicos e operacionais que contribuem para a deterioração de tubulações em sistemas de abastecimento de água (InfraGuide, 2003 - modificado).

Fatores que contribuem para a **deterioração** de tubulações em sistemas de abastecimento de água	Fatores físicos ou estáticos	Material do tubo	
		Diâmetro do tubo	
		Espessura da parede do tubo	
		Profundidade do tubo	
		Tipo de junta	
		Revestimento do tubo	
		Proteção catódica	
		Aterro da vala do tubo	
		Condições de fabricação	
		Condições de transporte e de armazenamento prévio à instalação	
		Condições de instalação frente aos requisitos de assentamento do tubo	
	Fatores dinâmicos	Ambientais	Clima
			Temperatura
			Terremotos
			Deslizamentos de terra
			Umidade do solo
			Águas subterrâneas
			Resistência elétrica
			Tipo do solo
		Cargas externas	Cargas de tráfego
		Idade da tubulação	
	Fatores operacionais	Perdas de água	
		Práticas de operação	
		Práticas de manutenção	
		Intermitência do abastecimento	
		Qualidade da água	
		Pressões operacionais de serviço	
		Velocidade do fluxo	

3.1.2. Consequências da falha estrutural

Resta elencar a gama de potenciais consequências oriundas falha estrutural que não estão apenas ligadas à infraestrutura dos sistemas de abastecimento de água, mas também às outras infraestruturas existentes em um centro urbano. As consequências podem incluir perturbações socioeconômicas e impactos ambientais, por exemplo. Portanto, ao avaliar o risco associado a um evento específico, várias dimensões precisam ser consideradas. A Tabela 3.4, adaptada de Almeida *et al*. (2011), ilustra critérios que podem ser utilizados na avaliação das consequências do rompimento de tubulações.

Tabela 3.4 – Dimensões e consequências do rompimento de tubulações (Almeida *et al*., 2011 - modificado)

Dimensão	Tipos de variáveis para expressar valor relativo em cada classe
Saúde e segurança	Número e gravidade das lesões; Número e gravidade das pessoas afetadas; e, Número de pessoas afetadas permanentemente (mortalidade e invalidez).
Financeira	Valor monetário (deve ser em função do porte financeiro, por exemplo, em relação ao faturamento anual).
Continuidade do serviço	Duração da interrupção do serviço prestado; e, Perfil de clientes afetados.
Impactos ambientais	Gravidade (expressa, por exemplo, como o tempo esperado de recuperação); Extensão (volume ou duração do evento); e, Vulnerabilidade (áreas protegidas ou zonas estratégicas, como áreas de drenagem à captações de água)
Impacto funcional no sistema	Medidas de desempenho (por exemplo, continuidade no abastecimento)
Reputação e imagem	Número de reclamações; e, Número de vezes em que a operadora é vinculada a notícias negativas em meios de comunicação
Continuidade dos negócios	Danos materiais; Capacidade de serviços; Perdas de água; e, Tempo de recuperação do sistema.
Desenvolvimento de projetos	Efeitos sobre desvio de objetivos (por exemplo, escopo, cronograma, orçamento)

Observa-se que as dimensões potencialmente afetadas pelo rompimento de uma tubulação são variadas e geram consequências diretas e indiretas:

- Consequências diretas: danos a propriedades, danos à saúde humana, danos ambientais, perda de produção, custos de reparos, custos de limpeza e remediação; e,
- Consequências indiretas: litigações e violações contratuais, insatisfação de clientes, reações políticas, perda de mercado, multas e penalidades governamentais.

Sob o ponto de vista da continuidade dos negócios, as perdas de água nos sistemas de abastecimento possuem papel de destaque, uma vez que altas perdas de água indicam ineficiência do serviço prestado.

Portanto, o próximo item aborda questões relacionadas à falha estrutural das tubulações de sistemas de abastecimento e às perdas de água.

3.2. PERDAS EM SISTEMAS DE ABASTECIMENTO DE ÁGUA

As perdas de água nos sistemas de abastecimento de água brasileiros apuradas em 2019 constatam que 39,2% da água captada é perdida (SNIS, 2019), o que representa margem considerável de melhoria dos sistemas, buscando-se eficiência, economicidade e melhores condições para a prestação dos serviços em busca da universalização do acesso à água com qualidade e quantidades adequadas à população. A Figura 3.1 e a Figura 3.2 apresentam, respectivamente, os valores das perdas de água por companhia de saneamento em termos percentuais e em litros/ligação/dia.

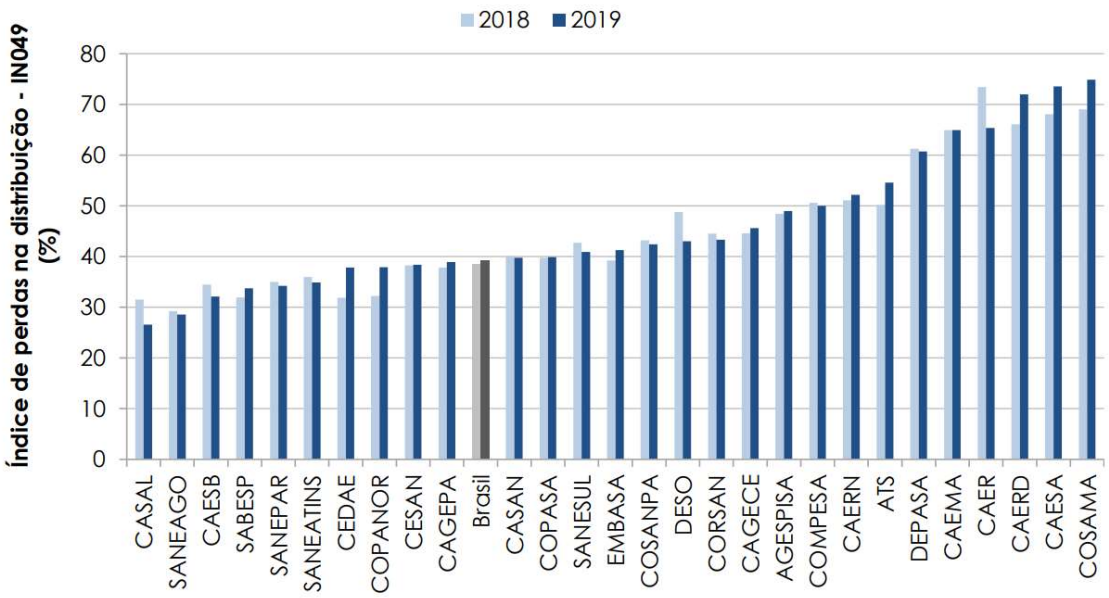


Figura 3.1 – Índice de perdas na distribuição dos prestadores de serviços de abrangência regional participantes do SNIS em 2018 e 2019, segundo prestador de serviços (fonte: SNIS, 2019)

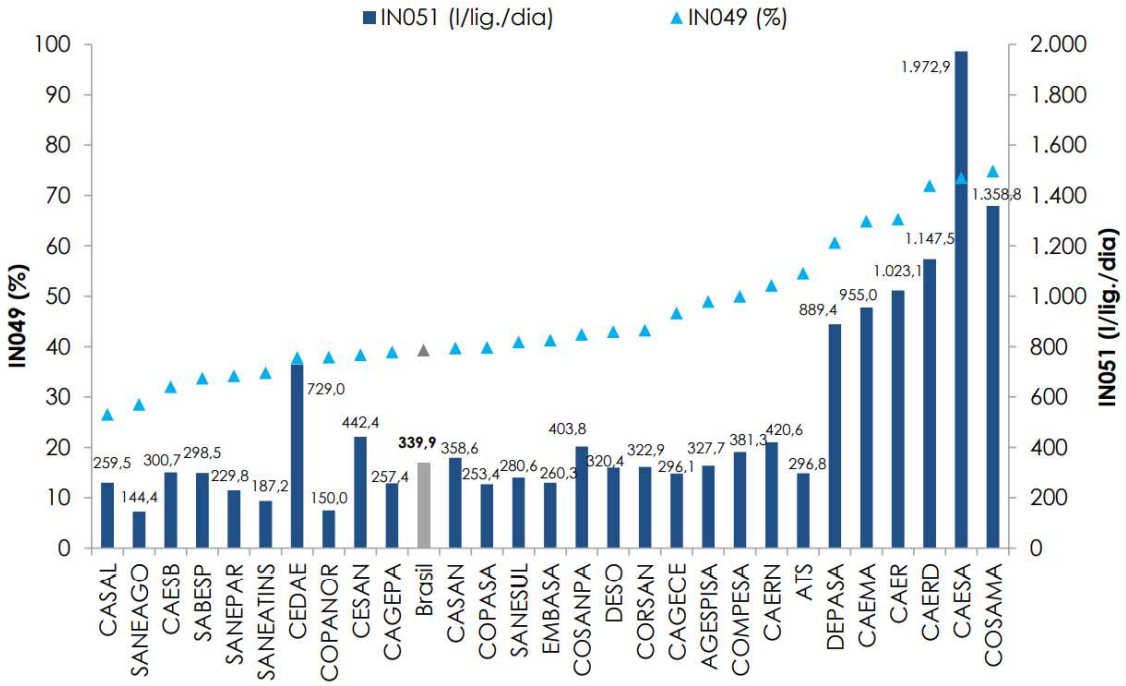


Figura 3.2 – Índice de perdas por ligação e na distribuição dos prestadores de serviços de abrangência regional participantes do SNIS em 2019, (fonte: SNIS, 2019)

Embora as perdas de água no Brasil, em termos gerais, estejam relativamente altas, cabe pontuar que o desafio da gestão das perdas de água é realidade em vários países. Liemberger e Wyatt (2018) realizaram uma atualização das estimativas globais sobre perdas de água publicadas em estudo realizado pelo Banco Mundial (Kingdom *et al*., 2006). Tais autores propõem que o volume perdido é de 126 bilhões m³/ano, cobrindo custos financeiros de 39 bilhões de dólares/ano; concluindo que a incitação frente às mudanças climáticas, expansão da população e a necessidade de cobertura adequada da prestação dos serviços são um desafio global.

3.2.1. Balanço Hídrico

O consagrado método de Balanço Hídrico para cálculo de perdas de água (Hirner e Lambert, 2000; Alegre *et al*., 2000) é o fundamento técnico da classificação das perdas. Essa abordagem compartimenta os volumes de água de entrada no sistema em dois grupos: Consumo Autorizado; e, Perdas de Água; com as suas devidas subdivisões, conforme apresentado no Quadro 3.2.

Quadro 3.3 – Componentes do Balanço Hídrico (Alegre *et al*., 2017 - modificado).

<table>
<tr><td rowspan="9">Volume de entrada no sistema
(m³/ano)</td><td rowspan="3">Consumo autorizado
(m³/ano)</td><td rowspan="2">Consumo autorizado faturado
(m³/ano)</td><td>Consumo medido faturado (incluindo água exportada)
(m³/ano)</td><td rowspan="2">Água faturada
(m³/ano)</td></tr>
<tr><td>Consumo medido não faturado
(m³/ano)</td></tr>
<tr><td rowspan="2">Consumo autorizado não faturado
(m³/ano)</td><td>Consumo medido não faturado (m³/ano)</td><td rowspan="7">Água não faturada
(m³/ano)</td></tr>
<tr><td>Consumo não medido não faturado (m³/ano)</td></tr>
<tr><td rowspan="6">Perdas de água
(m³/ano)</td><td>Perdas aparentes
(m³/ano)</td><td>Consumo não autorizado (m³/ano)</td></tr>
<tr><td></td><td>Imprecisões de medição
(m³/ano)</td></tr>
<tr><td rowspan="4">Perdas reais
(m³/ano)</td><td>Perdas reais em adutoras de água bruta e em processos de tratamento
(m³/ano)</td></tr>
<tr><td>Vazamentos em redes de distribuição ou em adutoras de água tratada</td></tr>
<tr><td>Vazamentos e extravasamentos em reservatórios de distribuição</td></tr>
<tr><td>Vazamentos nos ramais de ligação/serviço até o ponto de hidrometração</td></tr>
</table>

As definições abreviadas dos principais componentes do balanço hídrico IWA (Hirner e Lambert, 2000; Alegre *et al*., 2000) são as seguintes:

O volume de entrada do sistema é o volume anual de água fornecido a um sistema de abastecimento de água. Consumo Autorizado é o volume anual de água medida ou estimada consumida por clientes cadastrados, o próprio prestador do serviço de águas e outros usuários que estão implícita ou explicitamente autorizados a fazê-lo.

A Água Não Faturada, *Non-Revenue Water* (NRW), é a diferença entre o Volume de Entrada do Sistema e o Consumo Autorizado Faturado. A Água Não Faturada consiste em:

- Consumo autorizado não faturado; e,
- Perdas de água.

Segundo Hirner e Lambert (2000), as perdas de água são a diferença entre o volume de entrada do sistema e o consumo autorizado, e consistem em perdas aparentes e perdas reais. As perdas aparentes consistem em consumo não autorizado e todos os tipos de imprecisões de medição. Perdas reais são os volumes anuais perdidos por meio de vazamentos e rompimentos na rede de distribuição e nos ramais de ligação até o ponto de medição do cliente, além de extravasamento em reservatórios de distribuição do sistema.

3.2.2. Gestão de Perdas Reais

Embora seja natural o anseio pela eliminação das perdas, é salutar pontuar que as perdas podem ser minimizadas, mas não totalmente eliminadas. O objetivo da gestão das perdas reais é atingir o patamar de menor custo entre a quantidade de água perdida e o custo das atividades de controle, com o objetivo de maximizar o valor à sociedade. Este é o conceito de Nível Econômico de Perdas (ELL – *Economic Level of Leakage*). O investimento em redução ou controle de perdas de água deve ser precedido por análise econômica à luz dos indicadores de performance da rede de distribuição (Alegre *et al*., 2000), ponderando-se que a receita recuperada da água deve ser no máximo igual ao custo do controle ativo de perdas, em custos marginais (Gonçalves, 1998). Segundo Trow e Hall (1994) o nível ótimo de perdas baseia-se no nível de vazamentos em que o custo marginal do controle ativo de perdas é igual ao custo marginal do vazamento, condição em que não há viabilidade econômica para o investimento em controle de perdas.

Liemberger e Wyatt (2018) consideram que o valor de um metro cúbico de água não faturada para uma operadora do sistema depende de uma série de fatores, tais como:

- A razão entre as perdas aparentes (comerciais) e as perdas reais (físicas);
- A cobrança das taxas do serviço de esgotos (caso realizada em função do volume de água consumido, as perdas comerciais devem considerar as taxas do serviço de água e, também, o de esgotos);
- Caso exista demanda reprimida, as perdas físicas recuperadas podem ser revertidas em atendimento, devendo-se considerar o valor da tarifa média a ser faturada;
- Não havendo demanda reprimida, a redução das perdas físicas suprime custos variáveis de captação, tratamento e distribuição; e,

- Pleiteando-se a expansão de fontes adicionais de água para cobrir acréscimos de demanda, o investimento e os custos operacionais necessários para viabilizar tais iniciativas devem ser confrontados com a redução das perdas, dado que tais condições impactam o valor da água não faturada.

De posse de um diagnóstico preciso sobre as condições das perdas de água, conforme Balanço Hídrico, sucedido de análise econômico-financeira e de indicadores técnicos, conduzindo-se às análises sobre o Nível Econômico de Perdas, as atividades de gestão são orientadas por meio das ações que possuem resultados mais atrativos. No âmbito das perdas reais, conforme reiterado por Lambert *et al*. (2014), segundo entendimentos prestados pela Força Tarefa em Perdas de Água da IWA, são métodos eficazes para sua gestão e controle: o controle ativo de vazamentos; a gestão da pressão operacional da rede; a velocidade e qualidade do reparo; e, a gestão da infraestrutura.

As pressões operacionais nos sistemas de abastecimento de água influenciam de forma indiscutível a ocorrência de rompimentos e vazamentos, sendo o gerenciamento da pressão nos sistemas uma das formas mais eficazes de combate às perdas de água, amplamente abordada em diversos estudos (Trow e Tooms, 2014). A rapidez e a qualidade dos reparos realizados nos vazamentos detectados na rede permitem a redução do volume de perdas (Lambert, 1994), garantindo que os volumes perdidos sejam controlados.

Ainda em relação aos volumes perdidos, além da questão sobre o tempo decorrido para reparo após a identificação do vazamento, é imprescindível caracterizar os vazamentos quanto ao seu comportamento. Os vazamentos podem ser classificados em vazamentos inerentes, não visíveis e visíveis. O primeiro tipo é intrínseco à infraestrutura e de difícil remoção, pois, inclusive, seu reparo esbarra nos próprios custos para realização da correção, uma vez que o volume perdido é muito baixo. Os vazamentos não visíveis são aqueles que não afloram à superfície e, portanto, carecem da aplicação de técnicas para o Controle Ativo de Vazamentos dado que a rede precisa ser inspecionada para identificação e reparo. Os visíveis, por sua vez, afloram à superfície e são de rápida localização. Tais condições induzem à seguinte condição: vazamentos visíveis possuem alta vazão, mas afloram e são rapidamente reparados; vazamentos não visíveis tendem a perder maiores volumes pois o tempo de identificação é superior; e, vazamentos inerentes

possuem larga duração, mas o volume perdido é baixo devido as suas características. A Figura 3.3 apresenta uma ilustração modificada de GIZ (2011) pontuando a distinção entre vazamentos inerentes, não visíveis e visíveis em redes de distribuição de água.

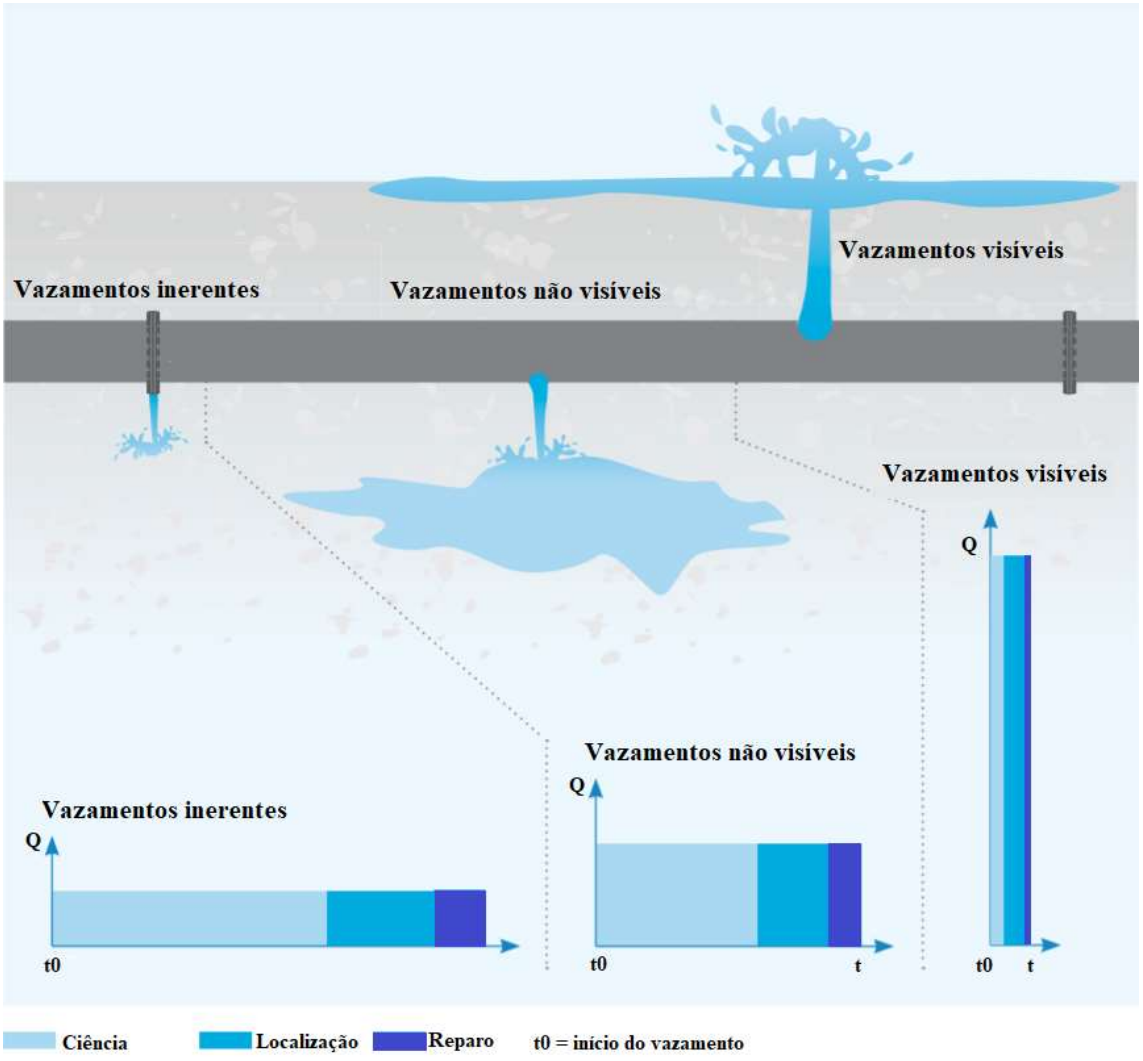


Figura 3.3 – Vazamentos inerentes, não visíveis e visíveis em redes de distribuição de água: volume de água perdido considerando os diferentes tipos de vazamento, tempos de ciência, localização e reparo do vazamento. (GIZ, 2011 - modificado).

As redes de distribuição de água são compostas por tubos, juntas, hidrantes, válvulas, conexões, curvas, entre outros acessórios. O ramal de ligação para atendimento às residências dos clientes é originado por meio de derivação às tubulações da rede de distribuição utilizando-se um elemento chamado de colar de tomada ou tê de serviço. Junto ao colar de tomada é realizada a perfuração ao tubo da rede e, então, a ligação de ramal de água é originada. Em geral, o ramal é composto por tubos e conexões até o

encontro do padrão de medição da concessionária, constituído, basicamente, pelo cavalete e o hidrômetro que realiza a medição do consumo realizado pelo cliente. A Figura 3.4 ilustra os componentes do ramal de ligação de água entre a derivação no tubo da rede por meio do colar de tomada ao padrão de medição com hidrômetro.

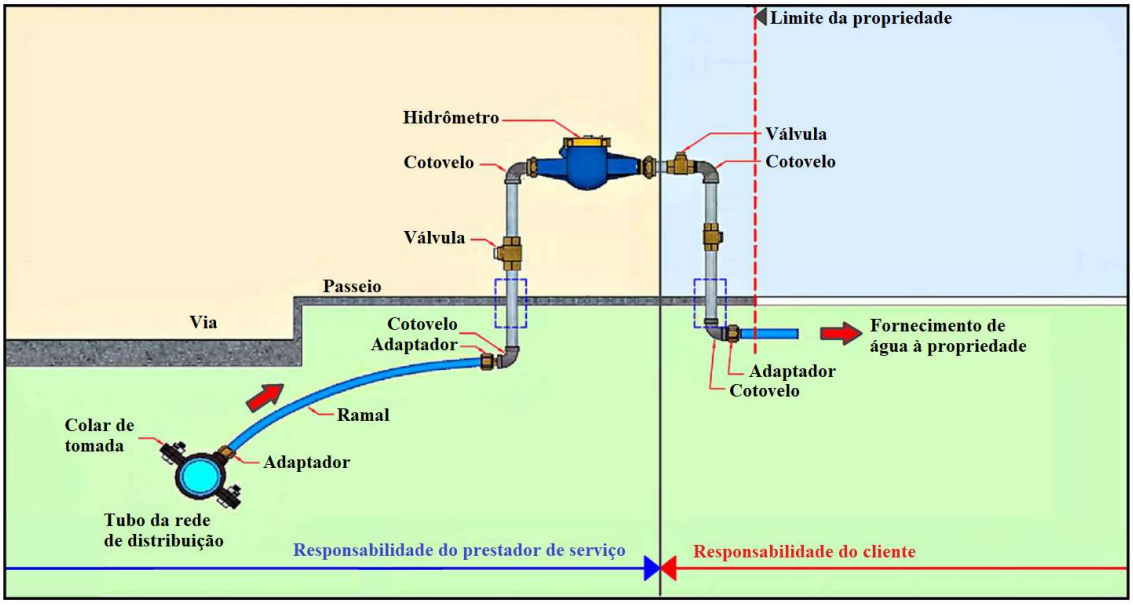


Figura 3.4 – Ilustração dos componentes constituintes do ramal de ligação de água entre a derivação da tubulação da rede ao padrão de medição e hidrômetro. (Taal Water District, 2016 - modificado).

As análises de componentes de perdas reais por meio do Balanço Hídrico contribuem com o diagnóstico dos fatores indutores das perdas, além de permitir a avaliação de como os componentes são influenciados pelas políticas da concessionária. O grande número de juntas e acessórios nos ramais de ligação entre o tubo da rede e o padrão de medição resulta em condições propícias para os vazamentos nesta parte da infraestrutura (Lambert *et al*., 2002). Segundos os autores, a frequência de novas quebras, por km de tubulação, é várias vezes maior nos ramais de ligação do que nos tubos da rede de distribuição propriamente dita. Os autores reiteram que embora as vazões instantâneas proporcionadas em rompimentos e vazamentos em tubos da rede de distribuição sejam superiores, os casos observados em ramais de ligação demandam mais tempo para serem reportados ou são de identificação mais complexa, sendo evidente que na maioria dos sistemas o maior volume de perdas reais anuais geralmente ocorre em ramais de ligação, conforme informações compiladas pelos autores sobre condições observadas em diversos operadores de sistemas de abastecimento de água por todo o mundo.

Tendo em vista a imprescindibilidade de gerir os vazamentos no sistema, em especial aqueles não visíveis e existentes em ramais de ligação, que representam parcela significativa dos volumes perdidos, o Controle Ativo de Vazamentos se apresenta como o conceito que trata sobre a gestão, identificação e reparo de vazamentos. Tal abordagem se traduz em uma estratégia proativa de redução das perdas de água por meio da detecção de vazamentos não visíveis, tipicamente por meio da pesquisa acústica dos ruídos gerados pelos vazamentos, cujo controle é possível por meio da ação de equipes especializadas e do monitoramento contínuo do sistema de abastecimento, sendo o uso de Distritos de Medição e Controle (DMC) uma técnica internacionalmente aceita e estabelecida para auxiliar neste processo (Charalambous *et al*., 2014). Segundo Serranito e Donnelly (2015), são meios de promover o Controle Ativo de Vazamentos: Inspeções Regulares; ou, Monitoramento e Controle Proativo das Perdas na Rede. A primeira abordagem consiste em campanhas extensas de varredura das redes, buscando-se identificar vazamentos não reportados ou não visíveis por meio do uso de instrumentos que identifiquem o vazamento. O segundo método consiste na adoção proativa do uso de instrumentação na rede e controle de indicadores de perdas à miúde para a rápida constatação de sinistros no sistema para sua pronta pesquisa, o que consiste em estratégia de maior êxito no combate às perdas, pois o ciclo de vida do vazamento tende a ser encurtado. Para ambas as abordagens, é salutar que o operador do sistema de águas possua ferramentas que possam permitir maior produtividade e assertividade no que tange identificar e reparar vazamentos.

Sob uma perspectiva de curto prazo, a gestão das perdas reais deve ser realizada para reduzir a duração dos vazamentos por meio da execução de intervenções ágeis para sanar vazamentos reportados, além de melhorar a qualidade do reparo executado. Para obter resultados a médio e longo prazo, reduzindo efetivamente as perdas na rede, é imprescindível promover ações no sentido de reduzir as pressões de serviço, gerir eficientemente os ativos da empresa, e executar o controle ativo de vazamentos (Serranito e Donnelly, 2015).

A implementação de um sistema integrado de gestão das perdas de água e gestão de ativos por uma concessionária ajuda a garantir o equilíbrio entre desempenho, custo e risco do ativo, contribuindo para uma política de manutenção, reposição ou renovação efetiva e sustentável (Alegre *et al*., 2012).

3.3. MODELOS HIDRÁULICOS DE REDES DE DISTRIBUIÇÃO DE ÁGUA

A compreensão completa do funcionamento de redes de distribuição de água é algo complexo, pois a infraestrutura encontra-se enterrada e foi construída, em geral, há décadas – antes dos avanços atuais em termos cadastrais proporcionados pelos Sistemas de Informações Geográficas (SIG). Soma-se a estas condições a impossibilidade de se instrumentar toda a rede para a realização direta do monitoramento das grandezas hidráulicas em toda a extensão da rede. Assim, devido às incertezas inerentes a este tipo de prestação de serviço, o uso de modelos hidráulicos das redes permite que o monitoramento de pontos do sistema possa embasar considerações e extrapolações ao funcionamento da rede, estimando-se velocidades e pressões em pontos não monitorados. Tal método permite que soluções técnicas possam ser avaliadas e complementadas por meio do estudo da dinâmica da rede.

Para sua utilização, os modelos de distribuição de água precisam ser submetidos a um processo de calibração, onde são ajustados os parâmetros e condições de contorno da rede, tais como rugosidade de tubos e demanda (Savic *et al*., 2009b). Tal processo é imprescindível para minimizar as diferenças entre os resultados de uma simulação e os dados obtidos em campo por monitoramento, tornando o modelo confiável (Walski, 1983).

Entretanto, é salutar esquadrinhar que a aplicação de modelos hidráulicos possui limitações e que tais condições versam sobre a qualidade e integridade dos dados de entrada nos modelos, além das condições e métodos de calibração utilizados. A incerteza dos resultados da modelagem de redes de água pode ser causada por muitos fatores, que podem ser classificados de acordo com Hutton *et al*. (2014) em incertezas estruturais, de medições ou de parâmetros. A incerteza estrutural está relacionada à representação do sistema real, como agregação ou esqueletização do modelo. A incerteza da medição diz respeito à incapacidade dos dispositivos de medição de capturar a variação temporal e espacial da demanda do consumidor e aos erros relacionados à própria medida. Outra fonte de incerteza está relacionada à presença de vazamentos na rede de distribuição, que tem sido amplamente estudada na literatura (Puust *et al*., 2010; Covelli *et al*., 2015; Covelli *et al*., 2016). De acordo com Kang e Lansey (2011), a rugosidade dos tubos e as demandas de água são os parâmetros de entrada mais incertos em um modelo hidráulico

porque não são mensuráveis diretamente dado que é impossível aferir a rugosidade real de cada tubo enterrado no sistema. Ademais, de acordo com Gouveia e Soares (2021), a forma como as demandas são distribuídas em uma rede de distribuição simulada pode afetar significativamente os resultados obtidos pelo processo de calibração, uma vez que a calibração de rugosidade perde efetividade caso as vazões distribuídas na rede estejam pouco precisas e distantes da realidade – baixa qualidade na alocação de demanda induz à baixa precisão de calibração, culminando em riscos à utilização dos modelos por divergirem significativamente das condições reais da rede.

Savic *et al*. (2009b) apresentam uma revisão da literatura sobre calibração de redes de distribuição de água, classificando os métodos de calibração em três categorias diferentes. A primeira trata de procedimentos iterativos e por tentativa e erro, em que os parâmetros desconhecidos são atualizados a cada iteração (Walski, 1983; Brave, 1988); culminando em uma taxa de convergência lenta. Em segundo lugar, apresenta-se métodos explícitos que são baseados na solução de um conjunto estendido de equações (Ormsbee e Wood, 1986), composto por equações iniciais e secundárias derivadas das medições disponíveis. A terceira categoria é composta por métodos implícitos ou inversos baseados em técnicas de otimização, buscando-se minimizar desvios entre os valores observados e simulados de pressão e vazão (Soares, 2003), considerando duas restrições: equação de energia e massa, implícitas na hidráulica do problema. Nesta linha, diversas abordagens diferentes têm sido propostas pela literatura, por exemplo, com base em algoritmos de objetivo único (Tabesh, 2011; Di Nardo *et al*., 2014; Do *et al*. 2016) ou multiobjectivo (Asadzadeh, 2011). Além disso, as variáveis de calibração consideradas têm uma ampla gama de parâmetros possíveis, como demanda nodal e rugosidade do tubo (Wu, 2002), ou estado da válvula e parâmetros de vazamento (Laucelli, 2011).

Meirelles *et al*. (2017) propôs um modelo baseado em uma rede neural artificial para prever a pressão nos nós da rede. Posteriormente, a calibração foi realizada usando uma Otimização por Enxame de Partículas para estimar a rugosidade dos tubos, minimizando a função objetivo descrita como a diferença entre a pressão simulada e a prevista. Do *et al*. (2017) propôs uma abordagem para estimar a demanda em tempo quase real em redes de água. Uma metodologia preditora é aplicada para prever a hidráulica da rede e, em seguida, um modelo baseado em filtro de partículas é usado para calibrar as demandas de água. Zhou *et al*. (2018) desenvolveram um sistema auto adaptativo baseado na técnica

de Filtro de Kalman para desenvolver uma calibração dupla da rugosidade do tubo e das demandas de água nodal em um sistema de distribuição de água.

Do *et al*. (2017) propôs uma abordagem para lidar com o problema de calibração usando várias execuções de um modelo de algoritmo genético, verificando-se que uma boa solução pode ser alcançada calculando-se a média dos resultados obtidos em várias execuções da simulação hidráulica. Uma abordagem semelhante foi proposta por Letting *et al*. (2017), que propôs uma abordagem baseada na otimização por enxame de partículas. Dada a natureza estocástica do problema de calibração, tanto Do *et al*. (2017) quanto Letting *et al*. (2017) fizeram várias execuções de seus algoritmos de otimização e usaram a média das soluções como um resultado mais preciso.

Além do procedimento de calibração, a abordagem da modelagem hidráulica desempenha um papel crucial na precisão dos resultados. Na literatura, a maioria dos trabalhos é baseada no Epanet 2.0 (Rossman, 2000) e suas soluções hidráulicas (Di Nardo *et al*., 2014; Do *et al*., 2016; Letting *et al*., 2017; Sophocleous *et al*., 2017). O método numérico adotado nesse programa é baseado no algoritmo de Todini e Pilati (1988), que propôs uma solução direta das equações de conservação de massa nos nós e conservação de energia ao longo das tubulações das redes de água. A solução é garantida pela convergência do sistema de equações (Collins *et al*., 1978), mas como o problema é parcialmente não linear, uma linearização é realizada e obtida através da técnica gradiente de Newton-Raphson. O sistema linear resultante é resolvido com um procedimento iterativo para encontrar as pressões nodais e as vazões no tubo, chamado de Algoritmo de Gradiente Global.

3.4. MODELOS PREDITIVOS E MODELOS DE APRENDIZADO DE MÁQUINA

Os modelos preditivos podem ser classificados em simplistas, físicos, estatísticos e modelos baseados aprendizado de máquina (Wu e Liu, 2017; Snider e McBean, 2020). Modelos simplistas hierarquizam a probabilidade de falha dos tubos por meio de um único parâmetro, limitando-se apenas a avaliações incipientes sobre a dinâmica de falha nos tubos. Os modelos físicos, a fim de prever a propensão dos tubos à quebra, analisam a carga aplicada aos tubos e sua capacidade de resistir a ela, à luz da corrosão que surge

nas paredes internas e externas, por exemplo. Apesar de sua precisão, os modelos físicos comparados com outras abordagens demandam significativo volume de dados, requerem recursos econômicos expressivos e testes em escala real para a avaliação física e pormenorizada dos processos de deterioração da tubulação (Rajani e Kleiner, 2001; Savic *et al*., 2009a).

Modelos estatísticos podem ser classificados como determinísticos ou probabilísticos. O primeiro tipo opera por meio da definição específica de relações entre variáveis de entrada e de saída; o segundo, probabilístico, predita a probabilidade de ocorrência de um determinado evento. Quanto à aplicação em estudos de falha em redes de água, os modelos estatísticos usam dados históricos de quebra disponíveis para identificar padrões de falha de tubulação, correlacionando parâmetros de entrada com o evento de quebra (Rajani e Kleiner, 2001b; Scheidegger *et al*., 2015); sendo capazes de vincular os padrões de falha às variáveis descritivas do tubo; por exemplo, diâmetro, idade e comprimento (Kakoudakis *et al*., 2017). Comparativamente aos modelos físicos, os modelos estatísticos demandam menores recursos financeiros e são mais ágeis, portanto, são modelos amplamente utilizados para auxiliar estudos sobre falha em redes de distribuição de água (Wilson *et al*., 2015). Segundo Snider e McBean (2018), uma vez que muitas concessionárias têm registros extensos de falhas de tubulação, a aplicação de modelos estatísticos a redes inteiras de tubulação é viável (comparativamente aos modelos físicos); por isso, segundo os autores, muitos utilitários tendem a favorecer modelos estatísticos em vez de modelos físicos.

Aprendizado de Máquina é a base técnica da mineração de dados, sendo um subtipo de Inteligência Artificial, que emergiu da ciência da computação focada no estudo de algoritmos (Bishop, 2006). Constitui-se de métodos amplamente utilizados para encontrar, estudar e descrever padrões em dados, sendo a regressão e a classificação os tipos de aplicação mais frequentes (Witten *et al*. 2016). Além das questões quanto a seleção e aplicação de métodos de processamento, recomenda-se atenção à qualidade dos dados, pois muitos problemas podem surgir, tais como dados insuficientes, dados de baixa qualidade, dados incorretos, dados ausentes, dados irrelevantes, valores de dados duplicados e assim por diante (Mohri, 2018).

Além dos modelos tradicionais de Aprendizado de Máquina, pode-se citar o subtipo *Ensemble Learning Models*, que operam combinando diferentes técnicas para a melhoria de resultados (Optiz e Maclin, 1999; Polikar, 2006; Rokach, 2010), além da hiper parametrização para configurações com melhor desempenho (Bergstra e Bergio, 2012, Claesen e De Moor, 2015).

Em linhas gerais, segundo Mukhiya e Ahmed (2020), a aplicação de conceitos e métodos da Ciência de Dados, e consequentemente as técnicas de Aprendizado de Máquina, requerem várias fases de análise de dados, incluindo: análise de requisitos de dados; coleta de dados; processamento de dados; limpeza de dados; análise exploratória de dados; modelagem e algoritmos; produção de informações; e, comunicação. Especificadamente sobre as etapas para a utilização da Aprendizado de Máquina, Campesato (2020a) propõe as seguintes tarefas:

- Obter o conjunto de dados;
- Realizar a limpeza da base de dados (*data cleaning* ou *data cleansing*). Este processo consiste em detectar, corrigir ou até remover registros corrompidos, incompletos, incorretos, duplicados, irrelevantes, ou imprecisos, de um banco de dados;
- Selecionar variáveis para aplicação dos modelos;
- Aplicar redução de dimensionalidade;
- Selecionar algoritmo;
- Selecionar dados de treinamento e dados de teste;
- Treinar o modelo;
- Testar o modelo;
- Ajustar o modelo; e,
- Obter métricas do modelo.

Muitas vezes as bases de dados para aplicações de Aprendizado de Máquina não estão consistidas, limpas e prontas para aplicação. Segundo Campesato (2020a), a etapa de limpeza dos dados é adequada e frequentemente utilizada. A fase de preparação de dados envolve 1) examinar as linhas para garantir que contêm dados válidos (o que pode exigir conhecimento técnico e específico do domínio) e 2) examinar as colunas (seleção ou extração de recursos) para determinar a possibilidade de retenção apenas das colunas mais

importantes. Segundo Unpingco (2016), a limpeza de dados é uma tarefa fundamental em ciência de dados, pois registros errados ou de baixa qualidade podem ser prejudiciais aos processos e análises, induzindo a interpretações equivocadas ou reduzindo a acurácia dos modelos. As tarefas que perfazem a limpeza dos dados podem envolver: correção de dados errados, decisão sobre dados faltantes, decisão sobre dados duplicados e decisão sobre *outliers*, serviços que podem ser necessários para melhorar a consistência da base de dados utilizada, fundamental para melhorar a resposta obtida pelos modelos (Unpingco, 2016).

3.4.1. Classificação dos modelos de Aprendizado de Máquina

Segundo Mohri (2018), os principais métodos de Aprendizado de Máquina podem ser classificados, quanto tipo de aprendizado/treinamento, em:

- Aprendizagem supervisionada: o modelo recebe um conjunto de dados com exemplos rotulados para treinamento e, então, faz previsões para os dados não vistos. Este é o cenário mais comum associado a problemas de classificação e regressão.
- Aprendizagem não supervisionada: o modelo recebe exclusivamente dados de treinamento não rotulados e presta classificação para todos os pontos não vistos. Como, em geral, nenhum exemplo rotulado está disponível nesse ambiente, pode ser difícil avaliar quantitativamente o desempenho do modelo. O agrupamento, *clustering* em inglês, e a redução da dimensionalidade são exemplos de problemas de aprendizagem não supervisionados.
- Aprendizagem semissupervisionada: o modelo recebe uma amostra de treinamento que consiste em dados marcados e não marcados e faz projeções para todos os pontos não vistos. O aprendizado semissupervisionado é comum em ambientes onde os dados não rotulados são facilmente acessíveis, mas a obtenção dos rótulos é cara. Vários tipos de problemas que surgem em aplicativos, incluindo classificação, regressão ou tarefas de classificação, podem ser enquadrados como instâncias de aprendizagem semissupervisionada. A ideia por trás deste método é que a distribuição de dados não rotulados acessíveis ao modelo possa ajudá-lo a obter um desempenho melhor do que no ambiente totalmente supervisionado.

Quanto ao tipo, os modelos de Aprendizado de Máquina podem ser classificados (Campesato, 2020a):

- Regressão. A regressão é uma técnica de aprendizado supervisionado para prever quantidades numéricas. Um exemplo de uma tarefa de regressão é prever o valor de uma determinada ação. Quantificar e predizer valores futuros é uma tarefa diferente de prever qualitativamente se o valor de uma determinada ação aumentará ou diminuirá amanhã (ou algum outro período de tempo futuro). São exemplos de modelos regressores: Regressão Linear, Regressão Ridge, Regressão Lasso, Regressão Polynomial e Regressão Linear Bayesiana.
- Classificação. A classificação também é uma técnica de aprendizado supervisionado, todavia, aplicada à predição de categorias. Um exemplo típico de tarefa de classificação é detecção da ocorrência de fraudes ou classificação de imagens. São exemplos de modelos classificadores: *Decision Tree* (árvore única), *Randon Forest* (árvores múltiplas), *K-nearest Neighbor* (KNN), Regressão Logística (apesar do nome), Naive Bayes, *Support Vector Machine* (SVM). Alguns algoritmos de aprendizado de máquina (como SVMs, *Random Forest* e kNN) oferecem suporte à regressão, bem como à classificação (Unpingco, 2016).
- Clusterização. É uma técnica de aprendizagem não supervisionada para agrupar dados por semelhança. Os algoritmos de clusterização colocam pontos de dados em *clusters* diferentes sem conhecer a natureza dos pontos de dados. Depois que os dados foram separados em diferentes *clusters*, é possível aplicar algoritmos como o SVM para realizar classificação. São modelos de clusterização: *k-Means*, *Meanshift* e *Hierarchical Cluster Analysis*.

Existe uma ampla gama de modelos de Aprendizado de Máquina, com diferentes métodos de processamento e custo computacional. Em termos de modelos aplicados à classificação, são alguns dos modelos disponíveis para processamento: *Linear* e *Radial SVM* (Cortes e Vapnik, 1995; Drucker *et al.*, 1997; Andrews *et al.*, 2003; Mohri, 2018), *Logistic Regression* (McCullagh e Nelder, 1989), *KNN* (Zhou, 2004; Jiang *et al.*, 2014; Campesato, 2020a), *Decision Tree* (Quinlan, 1986; Rokach e Maimon, 2008), *Naive Bayes* (Hand e Yu, 2001; Russel e Norvig, 2003), *Random Forest* (Breiman, 2001), *bagged KNN* (Gul *et al*., 2018), *bagged Decision Tree* (Kotsiantis *et al*., 2005), *AdaBoost* (Schapire, 2013), *Gradient Boosting* (Friedman, 2001), e *XGBoost* (Chen e Guestrin, 2016).

3.4.2. Análise Exploratória de Dados

A Análise Exploratória de Dados é um processo aplicado para examinar inicialmente o conjunto de dados disponível. Objetiva auxiliar o descobrimento de padrões, a detecção de anomalias, o teste de hipóteses e a verificação de suposições por meio do uso da estatística e de representações gráficas dos dados (Mukhia e Ahmed, 2020), sendo a biblioteca Pandas, em Python, amplamente utilizada em Ciência da Dados (Harrison e Petrou, 2020) para fins de: avaliação da distribuição dos dados, tratamento de valores ausentes no conjunto de dados, tratamento de outliers, remoção de registros duplicados, codificação de variáveis categóricas, normalização, entre outros processos.

3.4.3. Redução de Dimensionalidade

A Redução de Dimensionalidade é a tarefa de sintetizar e simplificar as diferentes dimensões em um conjunto de dados, preservando características relevantes (Kramer, 2013). Para a redução da quantidade de atributos em um conjunto de dados (Campesato, 2020b), são métodos relevantes não supervisionados: Análise do Componente Principal (ACP), em inglês *Principal Component Analysis* (PCA); *t-distributed stochastic neighbor embedding* (t-SNE); e, *Truncated Singular Value Decomposition* (SVD).

Segundo Campesato (2020b), os componentes principais são novos componentes criados por meio de combinações lineares das variáveis iniciais em um conjunto de dados, utilizando a variância como uma medida de informação: quanto maior a variância, mais importante é o componente. As principais vantagens do uso da Redução de Dimensionalidade são (Campesato, 2020a): o tempo de computação é reduzido devido o conjunto de dado processado requerer menos recursos; e, permite-se a representação gráfica dos componentes.

Um conjunto de dados com quatro ou mais componentes não permite visualização. Todavia, é possível apresentar visualmente os dados em grupos de três atributos, o que pode colaborar com a obtenção de *insights* sobre os dados. A Redução de Dimensionalidade permite que conjuntos com 3 ou mais atributos (e, portanto, 3 ou mais dimensões) possam ser representados em duas dimensões. A Figura 3.5 ilustra como a transformação de alta dimensionalidade para baixa dimensionalidade em um exemplo de conjunto de dados com três dimensões.

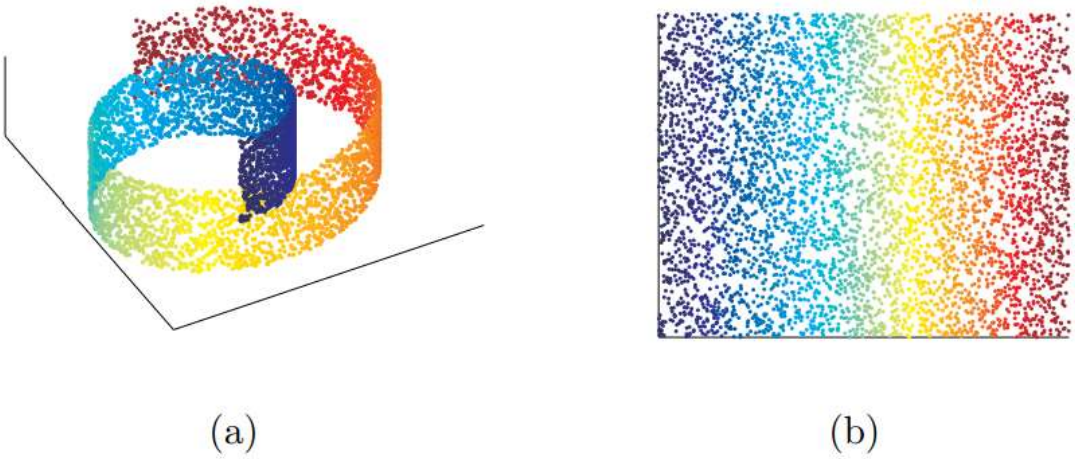


Figura 3.5 – Redução de Dimensionalidade. (a) representação de alta dimensão. (b) representação dimensional inferior. (Mohri *et al*., 2018).

3.4.4. Ensemble Learning Models

É comum ao processo decisório humano a consulta a outros indivíduos para ponderar diferentes percepções sobre uma determinada pauta que demande interpretação e decisão. Esse tipo de condição também é aplicado em Modelos de Aprendizado de Máquina. Por meio do conceito de *Ensemble Learning Models*, a arquitetura de solução absorve contribuições oriundas de diferentes modelos e os combina, a fim de se obter um classificador que supere a todos os demais individualmente (Rokach, 2019). Segundo Kumar e Jain (2020), são métodos de *Ensemble*: *Boosting*, *AdaBoosting*, *XGBoost* e *Stacking*.

Segundo Zhang e Ma (2014), o objetivo da arquitetura *Ensemble* é criar vários classificadores com tendências relativamente fixas (ou semelhantes) e, em seguida, combinar seus resultados para reduzir a variância. De acordo com os autores, erros de classificação são compostos por duas componentes: o viés (precisão do classificador); e, a variância (precisão quando treinado em diferentes conjuntos de treinamento); sendo que tais componentes possuem, frequentemente, uma relação de troca ou benefício/custo. A Figura 3.6 ilustra a redução da variância por meio do aprendizado conjunto. Em complemento, a Figura 3.7 ilustra um exemplo de etapas em *AdaBoost* com hiperplanos alinhados ao eixo como classificador de base, compilando limites de decisão em cada rodada de reforço até a composição da classificação robusta e final, construída por meio da combinação de classificadores fracos anteriores. Ainda segundo Zhang e Ma (2014),

o fato de que os resultados obtidos por modelos *Ensemble Learning Models* possam reduzir a probabilidade de escolha de um classificador de baixo desempenho, não necessariamente garantem melhor performance que modelos de soluções individuais. Portanto, ao longo do processo de investigação sobre qual modelo de aprendizado de máquina aplicar, é salutar avaliar o desempenho de diferentes abordagens para a definição da ferramenta a ser aplicada para subsidiar o processo decisório.

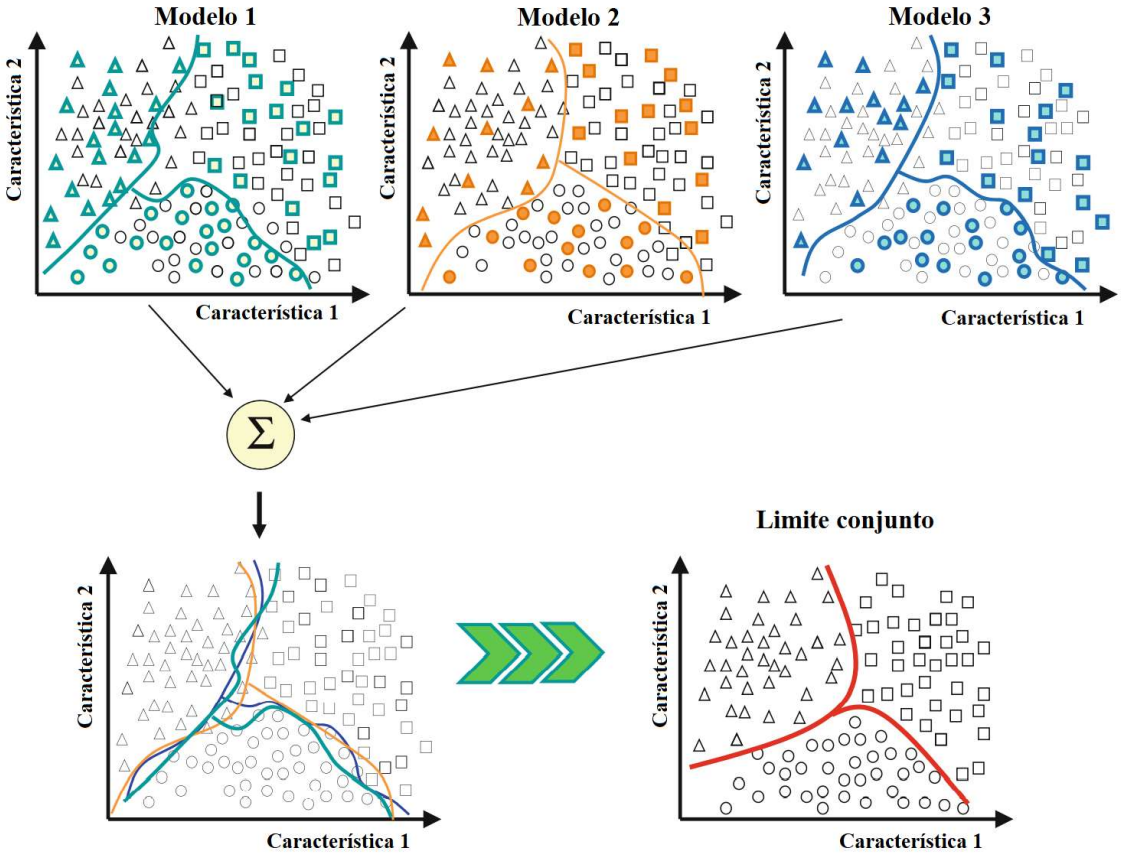


Figura 3.6 – Redução da variabilidade por meio de sistemas conjuntos (Zhang e Ma, 2014 – modificado).

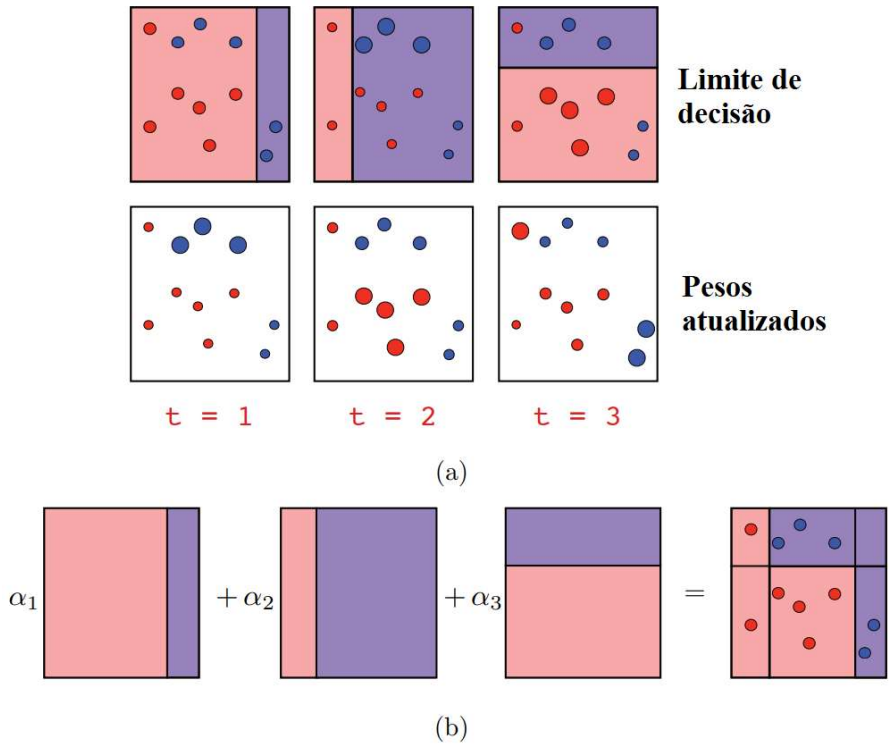


Figura 3.7 – Exemplo de *AdaBoost* com hiperplanos alinhados ao eixo como classificadores de base. (a) A linha superior mostra os limites de decisão em cada rodada de reforço. A linha inferior mostra como os pesos são atualizados em cada rodada. (b) Visualização do classificador final, construído como uma combinação de classificadores. (Mohri *et al*., 2018 – modificado).

3.4.5. Validação Cruzada

A Validação Cruzada é uma técnica utilizada para garantir a robustez do modelo, evitando-se *overfitting* quanto à base de dados (Mohri *et al*., 2018). Na metodologia de modelagem comum, um modelo é desenvolvido em dados de treinamento e avaliado em dados de teste. Em alguns casos extremos, os dados de treinamento e de teste podem não ter sido selecionados de forma homogênea, conduzindo os modelos a situações diferentes quanto ao treinamento e ao teste, o que prejudica o desempenho do modelo.

No método de Validação Cruzada, os dados são divididos em k partes iguais (k pastas, k partes ou, em inglês, *K-fold*); sendo que cada uma destas partes será subdividida originando uma amostra para treinamento e outra para teste, que permitirá a avaliação de desempenho do modelo (Unpingco, 2016; Dangeti, 2017). Este processo pode ser repetido quantas vezes o modelador definir, obtendo-se a estimativa geral de erro por meio da média das estimativas de erro (Russel, 2018).

Segundo Campesato (2020), o objetivo da validação cruzada é testar um modelo com conjuntos de teste não sobrepostos, procedendo-se da seguinte forma: etapa 1, os dados são subdivididos em k conjuntos de tamanhos iguais; etapa 2, seleciona-se um subconjunto para teste, utilizando-se os demais para treinamento; e, etapa 3, repete-se a etapa 2 para cada um dos subconjuntos restantes, até o uso de cada conjunto na condição de teste. Por exemplo, em uma validação cruzada quíntupla, os dados são divididos em cinco partes, posteriormente treinados em quatro partes dos dados e testados na quinta parte. Este processo será executado cinco vezes, a fim de cobrir todos os pontos dos dados, conforme ilustrado na Figura 3.8. Por fim, o erro calculado é a média de todos os erros.

	Dados				
1ª iteração	Teste	Treinamento	Treinamento	Treinamento	Treinamento
2ª iteração	Treinamento	Teste	Treinamento	Treinamento	Treinamento
3ª iteração	Treinamento	Treinamento	Teste	Treinamento	Treinamento
4ª iteração	Treinamento	Treinamento	Treinamento	Teste	Treinamento
5ª iteração	Treinamento	Treinamento	Treinamento	Treinamento	Teste

Figura 3.8 – Estrutura de Validação Cruzada (Dangeti, 2017 - modificado).

3.5. AVALIAÇÃO DE DESEMPENHO EM MODELOS DE APRENDIZADO DE MÁQUINA

O desempenho de métodos de Aprendizado de Máquina por Classificação pode ser avaliado usando-se Curvas de Precisão, Matrizes de Confusão e Curva Característica de Operação do Receptor (Curva COR), ou em inglês, *Receiver Operating Characteristic Curve* (*ROC Curve*).

A precisão é estimada como a fração das previsões corretas em relação às previsões totais, conforme Equação (3.1). A Matriz de Confusão é um dos métodos mais clássicos para avaliar e visualizar o comportamento da resposta de modelos de aprendizado de máquina supervisionados (Davis e Goadrich, 2006; Powers, 2011), apresentando o grau de confusão do algoritmo quanto à classificação dos dados processados (James *et al*., 2013). A Matriz de Confusão, apresentada na Tabela 3.5, apresenta informações sobre o desempenho de modelos de classificação, quantificando os resultados de acordo com as

previsões obtidas e observações disponíveis. Segundo Davies (2018), a classificação positiva e correta é representada como Verdadeiro Positivo (VP), enquanto a classificação negativa e correta é representada por Verdadeiro Negativo (VN). Classificações incorretas, por sua vez, são descritas por Falso Negativo (FN) ou por Falso Positivo (FP); o primeiro caso aborda predição negativa e dado amostral positivo, o segundo termo é utilizado para agrupar as predições positivas em relação a dados amostrais negativos.

$$Acurácia = \frac{(VP + VN)}{(VP + VN + FP + FN)} \tag{3.1}$$

Tabela 3.5 – Matriz de Confusão para classificações binárias (Fawcett, 2006 - modificado).

		Condição observada		
		Não	Sim	*Recall*
Condição predita	Não	Verdadeiro Negativo (VN)	Falso Negativo (FN)	VN/N
	Sim	Falso Positivo (FP)	Verdadeiro Positivo (VP)	VP/P
	Precisão	VN/(VN + FN)	VP/(VP + FP)	
		Total negativo	Total positivo	

Outras métricas, tais como: Taxa Verdadeiro Positivo (TVP), Taxa Verdadeiro Negativo (TVN), Taxa Falso Positivo (TFP), Taxa Falso Negativo (TFN) e a medida-F (*F-measure*), podem ser usadas para avaliar a capacidade preditiva de modelos. O TVP, ou *recall*, mede a porcentagem de previsões corretas feitas a partir da classe de interesse (que podem ser os tubos com falha, por exemplo). O TVN dá a porcentagem de classificação correta da outra classe (ou seja, os tubos que não falharam). Da mesma forma, o TFP apresenta a proporção de todos os negativos classificados incorretamente como positivos, e o TFN avalia os positivos classificados incorretamente como negativos. As taxas apresentadas estão relacionadas entre si através das Equações (3.2) e (3.3).

$$TFN + TVP = 1 \tag{3.2}$$

$$TFP + TVN = 1 \tag{3.3}$$

A medida-F compara o desempenho dos modelos em termos de *recall* e precisão (ou seja, uma medida de exatidão) usando um fator que controla sua importância relativa. A medida-F, precisão e *recall* tendem a 1 conforme o desempenho dos modelos aumenta.

Segundo Fawcetti (2006), as métricas TVP, TVN, TFP, TFN e medida-F são definidas conforme as Equações (3.4, (3.5, (3.6, (3.7 e (3.8, respectivamente.

$$TVP = Sensibilidade = Recall = \frac{VP}{VP + FN} \tag{3.4}$$

$$TVN = Especificidade = \frac{VN}{VN + FP} \tag{3.5}$$

$$TFP = 1 - Especificidade \tag{3.6}$$

$$TFN = 1 - Sensibilidade \tag{3.7}$$

$$Medida\text{-}F = \frac{2 * Precisão * Recall}{Precisão + Recall} \tag{3.8}$$

A Curva *ROC* é uma técnica útil para visualizar e selecionar o modelo mais adequado com base no desempenho (Russel, 2018). Essa curva é obtida traçando o TVP em função do TFP (Figura 3.9), considerando diferentes limiares de probabilidade para fazer previsões de classe. A Curva *ROC* é considerada confiável quando a curva está acima da linha de 45° (Fawcetti, 2006). A classificação perfeita é definida graficamente pela união de duas linhas, correspondendo a TVP igual a 1 e TFP igual a 1.

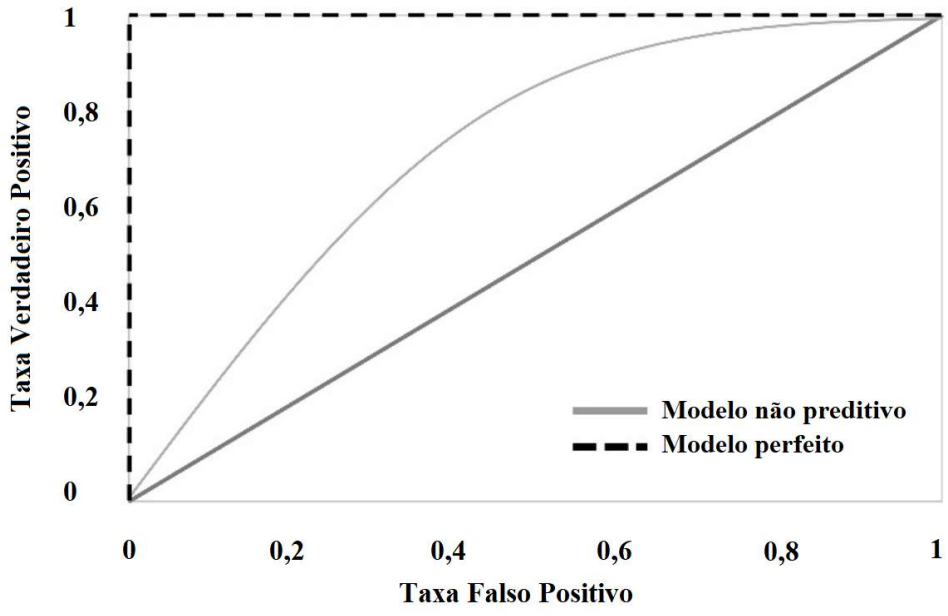


Figura 3.9 – Curva Característica de Operação do Receptor (Curva COR), ou em inglês, *Receiver Operating Characteristic Curve*, *ROC Curve* (Gönen, 2007 – modificado).

3.6. A APLICAÇÃO DE MODELOS PREDITIVOS À FALHA EM TUBULAÇÕES DE REDES DE DISTRIBUIÇÃO DE ÁGUA

A compreensão sobre os fatores intervenientes à falha em tubulações de redes de distribuição de água é fundamental para que modelos preditivos possam ser aplicados, subsidiando os processos de eficiência operacional e redução de perdas de água, ações estratégicas prioritárias para o uso sustentável dos recursos hídricos.

Ao longo dos últimos anos, métodos baseados em aprendizado de máquina têm sido aplicados à predição de falha em redes de distribuição de água, auxiliando o processo decisório quanto a identificação de tubos que demandam atenção urgente e aqueles que requerem substituição programada (Liu e Kleiner, 2014; Zhou e Chen 2018). Ademais, contribuem com a investigação de fatores intervenientes e variáveis explicativas do processo de quebra em tubulações, retroalimentando os processos de validação de dados e de aperfeiçoamento da aplicação dos modelos (Snider e McBean, 2020).

A seguir são apresentados, resumidamente, alguns estudos que investigaram fatores intervenientes e modelos preditores à falha em redes de distribuição de água baseados em modelos estatísticos ou modelos por aprendizado de máquina.

Achim *et al*. (2007) aplicaram Redes Neurais Artificiais aos dados de uma cidade australiana, Victoria, utilizando as seguintes variáveis de entrada: diâmetro, ano de construção, idade, comprimento e dados de localização das falhas em redes de distribuição de água, obtendo coeficiente de correlação máximo de 0,68 para prever falha em tubulações de ferro fundido. O autor sugere que o modelo pode ser aprimorado a partir da incorporação de outros fatores intervenientes à falha que possam corroborar com o fenômeno, tais como o tipo de método construtivo utilizado (não destrutivo *versus* destrutivo), experiência do agente executor da obra, cargas excessivas, e tipificação do tráfego existente de veículos; além de tecer comentários sobre a possibilidade de consideração de fatores temporais (como estações do ano e El-Niño), que alteram as condições ambientais de contorno climático e que poderiam interferir na incidência de quebras.

Yamijala *et al*. (2009) aplicaram modelos de Regressão Multivariada, Regressão Exponencial Multivariada, e Modelo Logístico Linear Generalizado para estimar a probabilidade de quebra em tubos, utilizando-se base de dados coletada entre 2000 e 2005 em uma companhia de saneamento no Texas, tais como: diâmetro do tubo, material do tubo, extensão do tubo, uso do solo, temperatura, umidade do solo, tipo do solo e a ocorrência de falhas. Resultados mostraram que o Modelo Logístico Linear Generalizado entregou melhores resultados que os demais. Todavia, apesar da inserção de um leque mais amplo de variáveis, os resultados obtidos pelos três modelos não indicaram boa aderência aos dados disponíveis.

Jafar *et al*. (2010) aplicaram Redes Neurais Artificiais para predizer falha em tubos de sistemas de abastecimento de água de uma cidade ao norte da França, utilizando-se dados coletados entre 1991 e 2014, perfazendo 424 falhas. O estudo contou com a utilização de variáveis físicas (material, comprimento, diâmetro, idade e espessura do tubo), variáveis ambientais (tipo do solo, localização do tubo na via) e operacionais dos tubos (pressões de serviço). Concluiu-se que a troca de 5% dos tubos da rede de distribuição estudada poderia evitar 51% das falhas, a troca de 10% das tubulações permitiria afastar 68% das falhas na rede de distribuição; tais achados sugeriram que o estudo e aplicação de modelos preditivos à falha podem ser utilizados para balizar o processo decisório para implementação de estratégias de reabilitação em redes de distribuição de água.

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, utilizando-se dados de Limassol, Cyprus, compondo 2000 eventos de falha observados entre 2002 e 2007. O autor reitera o potencial do uso da compreensão da falha nas tubulações, permitindo o aperfeiçoamento das estratégias de reparou ou substituição de tubulações.

Asnaashari *et al*. (2013) aplicaram Regressão e Redes Neurais Artificiais para predizer rompimento em redes de distribuição de água em Etobicoke, Ontário/Canadá, obtendo-se coeficientes de acurácia de 0,75 e 0,94, respectivamente. A base de dados consistida empregou informações sobre idade, diâmetro, material, comprimento tipo do solo e se há proteção catódica no tubo. Os resultados obtidos apresentaram significativa melhora da resposta do modelo por Redes Neurais Artificiais em detrimento do modelo por

Regressão, indicando que o método com melhor desempenho por meio da percepção de relações não lineares presentes no banco de dados. O autor sugere que o coeficiente de determinação obtido (R^2=0,94) indica que as Redes Neurais Artificiais apresentaram resultados promissores para sua utilização em processos de decisão para reabilitação da infraestrutura.

Francis *et al*. (2014) empregaram um modelo considerando fatores como material do tubo, diâmetro do tubo, idade do tubo, variáveis demográficas, pedológicas e climáticas para predizer quebra em tubulações por meio de Redes Bayesianas aplicadas. Utilizou-se de série de dados coletada entre 2010 e 2011 em cidade Norte Americana, contemplando 3686 registros de falha em redes de distribuição.

Shirzad *et al*. (2014) concluíram que o desempenho de *Support Vector Machine* se sobressaiu em relação a Redes Neurais Artificiais para predizer taxa de falha em tubulações de água em dois estudos de caso Iranianos. Ambos os modelos avaliados foram pautados em pressão hidráulica, diâmetro do tubo, extensão do tubo, idade do tubo, e profundidade.

Harvey *et al*. (2014) aplicaram Redes Neurais Artificiais para prever o tempo de falha de tubos em redes de distribuição de água considerando dados de Scarborough, Ontário/Canadá. A extensão de rede investigada foi de 1.021km, com idade média de tubos em 55 anos, sendo que 6,5% dos tubos possuíam, à época, mais de 100 anos de idade. A base histórica de registros utilizada contemplou falhas entre 1962 e 2005, utilizando-se dos seguintes critérios: ano de construção, material, comprimento, diâmetro, tipo de solo, ano de recomposição do revestimento interno (se realizado), ano da proteção catódica (se realizado) e histórico de falhas. Os modelos treinados exibiram coeficientes de determinação variando de 0,7 a 0,82, indicando potencial de uso às ações de gestão da infraestrutura.

Kabir *et al*. (2015) aplicaram modelos Bayesianos e Regressão para predizer a falha utilizando-se número de quebras anteriores, idade do tubo, diâmetro do tubo, extensão do tubo, resistência do solo, e corrosividade do solo, considerando dados de diferentes cidades canadenses. Os resultados apresentados indicam melhor performance do modelo Bayesiano. O autor recomenda a inserção de outras variáveis possivelmente

intervenientes à quebra em redes de abastecimento de água, tais como condições climáticas (precipitação, congelamento) e operacionais (velocidade e pressões de serviço), que podem melhorar o desempenho de modelos preditores à falha.

Aydogdu e Firat (2015) aplicaram *Support Vector Machine* para avaliação de fatores intervenientes e predição de falha em redes de distribuição de águas, utilizando-se dados coletados entre 2006 e 2012 em sistema de cidade turca. 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, inclusive citando que os registros históricos devem permitir a exclusão de falhas causadas por anomalias oriundas de serviços de terceiros, tais como companhias de telecomunicações, eletricidade, óleo e gás, etc), o que pode afetar a performance de modelos preditores, 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. Portanto, neste estudo, dos 21.000 registros de falha examinados, os autores compilaram para aplicação 5.111.

Demissie *et al*. (2017) desenvolveram um modelo de Redes Bayesianas Dinâmico para predizer quebra em redes de distribuição de água, utilizando aspectos estáticos, ambientais e operacionais como variáveis explicativas aplicada à infraestrutura de redes de Calgary. Empregou-se extensão de tubo, diâmetro do tubo, número de falhas prévias, tipo do ramal de serviço, índice de congelamento, índice de descongelamento, índice de chuva, e corrosão do solo. Os autores concluíram que os modelos propostos são robustos e válidos para ajudar os gerentes de serviços públicos e engenheiros a prever o total anual ou mensal de quebras de tubos para um único tubo ou sistema geral de tubos. Os autores sugerem que tais tipos de modelos podem ser integrados com ferramentas de gestão de ativos para a reabilitação de sistema de tubulações em curto ou longo prazo, contribuindo com o planejamento às operadoras.

Farmani *et al*. (2017) estudaram fatores estáticos e dinâmicos que impactam a condição de tubulações em sistemas de abastecimento de água no Reino Unido. O conjunto de dados foi dividido em grupos homogêneos se acordo com a similaridade de suas características, compondo-se grupos para a técnica de validação cruzada. A Regressão Polinomial Evolucionária foi aplicada para predizer o número de falhas baseando-se em

tipo de solo, diâmetro do tubo, e idade do tubo. Os resultados obtidos (R^2=0,89 para treinamento e R^2=0,83 para teste), indicam aplicabilidade à gestão das redes.

Parvizsedghy *et al*. (2017) aplicaram regressão à dados consistidos sobre falha em tubulações de sistemas de abastecimento de água em cidades canadenses. Os autores agruparam as variáveis utilizadas em três grupos: fatores físicos (material, idade, diâmetro e qualidade da instalação), fatores ambientais (profundidade das águas subterrâneas, tipo do solo e localização), e fatores operacionais (velocidade, histórico de quebra e qualidade da água). Os resultados obtidos apresentados pelo autor indicam bom potencial de predição das falhas, com R^2 superiores a 94% para os modelos desenvolvidos.

Kaushik *et al*. (2017) aplicaram um modelo de Regressão Logística para prever quebras de tubulação utilizando dados estáticos sobre diâmetro, comprimento, material, idade, profundidade, quantidade de falhas prévias; além de dados operacionais sobre pressões estáticas de serviço máximas e mínimas. O autor discorre sobre 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, o que foi o caso da área estudada pelo autor: o modelo hidráulico da rede de distribuição de água não se encontra total ou parcialmente construído e calibrado, ou pode estar desatualizado devido aos esforços necessários para manutenção das ferramentas de modelagem atualizadas. Segundo os autores, esta é provavelmente a razão pela qual os trabalhos existentes neste campo normalmente não consideram fatores operacionais dinâmicos (perda de carga, velocidade, vazão e pressão).

Sattar *et al*. (2017) sugeriram o uso de *Extreme Learning Machine* para prever a falha de adutoras com base em dados históricos de Scarborough, Grande Toronto, contemplando 1.000km de redes. O modelo foi treinado usando 9500 registros de falha com base no comprimento do tubo, material do tubo, método de proteção do tubo, e diâmetro do tubo. Os resultados obtidos pelo autor possuem coeficientes de determinação entre 0,67 e 0,82. Os principais fatores intervenientes foram a recorrência de quebras e o diâmetro do tubo; sendo que a proteção catódica representa um fator relevante na atenuação da taxa de falha

das tubulações. O autor recomenda a utilização de modelos preditores de falha em sistemas de abastecimento de água para auxílio das decisões em otimização de inspeções e manutenções programadas para lidar com o crescente custo do envelhecimento da infraestrutura, aprimorando os métodos de gestão dos serviços essenciais.

Gómez-Martínez *et al*. (2017) aplicaram Modelos Bayesianos à predição de falha em tubulações de redes de água considerando variáveis físicas (diâmetro, ano de instalação e material), ambientais (terreno, uso do solo e profundidade), e operacionais nos tubos (pressões máximas, mínimas e médias, além um índice de transientes, velocidades máximas, mínimas e médias). A base de dados utilizada refere-se a rede de Madrid, com registros coletados entre 2011 e 2014, que, conforme citado pelo autor, possui setorização e sistemas com modelos calibrados da rede de distribuição. Os resultados obtidos mostraram que o fator preponderante à falha foi o diâmetro, sendo imprescindível às análises. Surpreendentemente, o ano de instalação não se mostrou tão relevante; ao passo que a combinação diâmetro, características do terreno e grandezas hidráulicas (pressão e velocidade) compuseram boa predição, com acurácias superiores a 0,95. O autor conclui que as ferramentas estudas podem ser utilizadas para apoiar análises econômicas sobre estratégias entre reparo ou substituição de redes.

Snider e McBean (2018) aplicaram Redes Neurais Artificiais, *Decision Tree* e *XGBoost* para predizer falha em redes de distribuição de água a partir de dados consistidos em bases norte americanas. O melhor modelo preditor encontrado foi o *XGBoost* (com R^2=0,85), performando acurácias 1,2% e 25,9% superiores à Árvore de Decisão e Redes Neurais Artificiais, respectivamente. Foram utilizadas as seguintes variáveis: diâmetro, comprimento, ano de construção, revestimento, proteção catódica, tipo de solo, pressão operacional média, densidade de quebras, número de quebras prévias. A base de avaliação conteve 6.633 registros de falhas para 1.563 tubos. O ranqueamento da importância das variáveis utilizadas indica que o ano de instalação, a densidade de quebras, comprimento, quebras prévias, diâmetro e pressões operacionais médias são fatores preponderantes; as características do solo, por exemplo, apresentaram baixa relevância ao modelo. Os autores discorrem sobre a limitação de dados normalmente encontrada em bancos de dados reais, reiterando que aspectos como pressões de serviço e questões correlacionadas com transientes hidráulicos (tais como proximidade a estações de bombeamento e válvulas redutoras de pressão) podem contribuir com a investigação; por fim, cita que

variáveis ambientais como carga de tráfego de veículos e sazonalidade climática podem aperfeiçoar tais modelos.

Winkler *et al*. (2018) propuseram uma abordagem para prever falhas de tubos com base em técnicas impulsionadas (por exemplo, *RUSboost*, *Adaboost*, *Random Forest* e *Decision Tree*), com base nos atributos de tubos existentes e registros históricos de falhas em um sistema de médio porte. O algoritmo *RUSboost* teve o melhor desempenho para predizer falha em tubulações de sistemas de abastecimento de água, obtendo acurácia de 0.96. Tais resultados, concluem os autores, indicam que a aplicação de métodos de aprendizado de máquina são uma boa alternativa aos modelos estatísticos tradicionais de deterioração. As variáveis utilizadas foram agrupadas em: atributos físicos (falha prévia, idade, material, diâmetro, pressão nominal do tubo, comprimento, função operacional do tubo); atributos derivados de informações geográficas (número de conexões residenciais em uma mesma via, número de hidrantes em uma mesma via, número de válvulas no trecho, e número de válvulas em uma mesma via); e, atributos derivados de informações históricas (número total de falhas para o tubo, e número de falhas observadas em um tubo desde a sua substituição). As variáveis mais representativas observadas no estudo foram idade, material, comprimento e número de válvulas em uma mesma via.

Kumar *et al*. (2018) aplicaram diferentes técnicas de aprendizado de máquina à predição de falhas em redes de distribuição de água: *Decision Tree*, *Logistic Regression*, *AdaBoost*, *Random Forest*, e *Gradient Boosted Decision Tree*, utilizando características dos tubos (ano de instalação, diâmetro, tipo de solo, pressão de serviço, material, tráfego rodoviário, número de quebras prévias, e número de quebras próximas). Foram obtidas acurácias máximas de 62%.

Chen *et al*. (2019) investigaram a possibilidade de desenvolvimento de modelos preditores de falha em tubos de redes de distribuição em condições de ausência de dados confiáveis para aplicação. Tal abordagem se justifica, segundo o autor, devido à dificuldade de sistematização de dados em muitos prestadores de serviços de abastecimento de água, principalmente com sistemas informacionais legados e antigos. Portanto, a base de dados foi compilada utilizando-se quantidade de quebras em tubos agrupadas por década, dados demográficos (renda média, percentual de etnias em uma área) e informações ambientais (precipitação, escoamento superficial, tipo de solo, uso

do solo, entre outros). Os autores consideraram que avaliação de dados socioeconômicos e ambientais podem colaborar com a compreensão da dinâmica de expansão dos sistemas, além de fomentar o desenvolvimento de bases cadastrais avançadas para utilização.

Kerwin *et al*. (2019) aplicaram Redes Neurais Artificiais utilizando-se dados sobre quebras prévias em tubulações de água, diâmetro, ano de construção, pressão estática nos tubos e tipo de solo para predizer falha em grupos de ferro fundido e ferro dúctil, sugerindo que fatores relacionados com a consequência das falhas em tubulações podem agregar novas perspectivas ao uso de modelos preditores de falha, considerando aspectos qualitativos sobre risco e a consequência da falha, o que pode contribuir significativamente com a priorização de intervenções em redes de distribuição de água.

Alizadeh *et al*. (2019) avaliaram a quebra em redes de distribuição de água por meio de diferentes métodos: *Support Vector Regression*, Regressão Gaussiana, e Redes Neurais Artificiais. Os fatores considerados nas análises foram: material, diâmetro, idade, pressão hidráulica média, e comprimento do tubo; aplicados em estudo de caso com informações de Teerã, Irã. Os autores concluem reiterando que os recursos escassos para reabilitação da infraestrutura podem ser melhor aplicados considerando-se a identificação de zonas com maior risco de falha por meio de modelos preditores.

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

Robles-Velasco *et al*. (2020) compararam *Logistic Regression* e *Support Vector Machine* para prever falha em tubos de redes de distribuição de água com base no material do tubo, comprimento, idade, número de ramais no tubo, flutuação de pressão, e o número total de falhas. Os resultados obtidos mostraram que os métodos aplicados, Regressão Logística

e *Support Vector Machine*, obtiveram acurácias entre 0,75 e 0,80; além de permitir a análise de que a troca de 3% das tubulações da cidade estudada (Sevilha) preveniriam a falha em 30% dos casos.

Snider e McBean (2020) investigaram a aplicação de modelos preditores de falha em tubulações de água, avaliando-se 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. Portanto, os autores concluem ressaltando a necessidade de que tais estudos possuam componentes representativas sobre os fatores intervenientes à falha, de forma holística; embora muitos prestadores baseiem-se em 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 falha considerando 5 anos de dados em uma base de 2.000 registros.

Almheiri *et al*. (2020) aplicaram Redes Neurais Artificiais, Regressão *Ridge* e *Decision Tree* aplicadas à predição de falha em tubulações de redes de água, utilizando-se dados de Quebec, Canadá. A base de dados utilizada compôs-se de: material do tubo, diâmetro, comprimento, ano de instalação e quebras; perfazendo um histórico de 15 anos entre 1987 e 2001, totalizando 432km de redes de distribuição. Os resultados apresentados indicam o uso de *Decision Tree* como uma ferramenta com potencial de aplicação aos processos de gestão da infraestrutura. O autor cita que a limitação de dados e variáveis é crucial para a performance de modelos preditores, sendo relevante a composição de dados complementares aos processos dinâmicos e dependentes no tempo, tal como as características ambientais.

Giraldo-González e Rodrígues (2020) compararam a performance de modelos estatísticos de regressão (Linear, Poisson e Polinomial Evolucionária) e modelos de aprendizado de máquina (*Gradient-Boosted Trees*, *Support Vector Machine*, Bayes, e Redes Neurais Artificiais) aplicados à predição de falha em tubulações de redes de abastecimento de água. Os atributos utilizados foram agrupados em variáveis físicas (diâmetro, idade, comprimento), ambientais (umidade do solo, potencial de contração e expansão do solo, precipitação e uso do solo), e operacionais (número de válvulas no trecho, quantidade de hidrantes no trecho, e falhas prévias). Os modelos responderam com acurácia superior a 90%, indicando potencial de aplicação quanto a priorização de ações de reabilitação; com

margem para aperfeiçoamento por meio da inclusão de outras variáveis como pressões operacionais, sendo o autor.

Snider e McBean (2020b) aplicaram *Gradient-Boosted Tree* e *Survival Analysis* à predição de falhas em tubos de redes de água de uma Companhia prestadora de serviços Canadense, utilizando-se atributos como: idade, comprimento, proteção catódica, revestimento, tipo de solo, distrito, pressão de serviço e histórico de falhas. As correlações obtidas para *XGBoost* e *Survival Analysis* foram de 0,897 e 0,856, respectivamente. Ademais, os resultados obtidos pelo autor indicam uma tendência do *XGBoost* antecipar a previsão de falha, enquanto o erro obtido com *Survival Analysis* postergou a predição de falha, o que é citado como um efeito positivo, pois os operadores preocupados com a segurança e resiliência da prestação dos serviços de abastecimento podem aplicar tais métodos identificando e ranqueando locais para intervenção proativa.

A aplicação de modelos para predição de falhas em redes de distribuição de água tem sido um tema amplamente discutido no meio acadêmico nos últimos anos, conforme apresentado neste capítulo e sintetizado por meio da Tabela 3.6, a seguir.

Tabela 3.6 – Síntese da bibliografia pesquisada e estudada no contexto da presente pesquisa sobre falha em redes de água, sumarizando-se variáveis e métodos empregados.

Referência	Variáveis	Métodos
Achim *et al*. (2007)	Diâmetro, Idade, Comprimento	*Artificial Neural Network*
Yamijala *et al*. (2009)	Diâmetro, Idade, Material, Comprimento, Uso da terra, Tipo do solo, Umidade do solo, Temperatura	*Multilinear Regression Model, Multivariate Exponential Regression Model, Generalized Linear Model*
Jafar *et al*. (2010)	Diâmetro, Idade, Material, Comprimento, # de falhas, Pressão hidráulica, Tipo do solo, Espessura do tubo, Proteção catódica do tubo, Localização do tubo (calçada ou via)	*Artificial Neural Network*
Christodoulou (2011)	Diâmetro, Idade, Material, # de falhas, Tipo do incidente	*Possession Regression Model*
Asnaashari *et al*. (2013)	Diâmetro, Idade, Material, Comprimento, Tipo do solo, Proteção catódica do tubo	*Artificial Neural Network, Regression Model*
Francis *et al*. (2014)	Diâmetro, Idade, Material, Tipo do solo, Temperatura, Variáveis demográficas	*Bayesian Belief Networks*

Continuação da Tabela 3.6 – Síntese da bibliografia pesquisada e estudada no contexto da presente pesquisa sobre falha em redes de água, sumarizando-se variáveis e métodos empregados.

Harvey *et al*. (2014)	Diâmetro, Idade, Material, Comprimento, # de falhas, Tipo do solo, Proteção catódica do tubo	*Artificial Neural Network*
Kabir *et al*. (2015)	Diâmetro, Idade, Comprimento, # de falhas, Resistividade do solo, Corrosividade do solo	*Bayesian Model, Regression Model*
Aydogdu e Firay (2015)	Diâmetro, Idade, Comprimento, # de falhas	*Support Vector Machine*
Demissie *et al*. (2017)	Diâmetro, Idade, Comprimento, # de falhas, Corrosividade do solo, Espessura do tubo, Proteção catódica do tubo, Precipitação, Índice de congelamento, Índice de descongelamento, Tipo da conexão de serviço	*Dynamic Bayesian Network*
Farmani *et al*. (2017)	Diâmetro, Idade, Comprimento, Temperatura, Índice de congelamento	*Evolutionary Polynomial Regression*
Parvizsedghy *et al*. (2017)	Diâmetro, Idade, Tipo do solo, Tipo da superfície, Profundidade do lençol freático, Qualidade da instalação	*Regression Model*
Kaushik *et al*. (2017)	Diâmetro, Idade, Material, Comprimento, # de falhas, Pressão hidráulica, Velocidade, Profundidade do tubo	*Logistic Regression Model*
Sattar *et al*. (2017)	Diâmetro, Idade, Comprimento, # de falhas, Tipo do solo, Proteção catódica do tubo	*Extreme Machine Learning*
Gómez-Martínez *et al*. (2017)	Diâmetro, Idade, Material, Pressão hidráulica, Uso da terra, Tipo do solo, Profundidade do tubo	*Bayesian Model*
Snider e McBean (2018)	Diâmetro, Idade, Material, Comprimento, # de falhas, Pressão hidráulica, Tipo do solo, Proteção catódica do tubo, Densidade de quebras	*XGBoost, Artificial Neural Network, Decision Tree*
Winkler *et al*. (2018)	Diâmetro, Idade, Material, Comprimento, # de falhas, Quantidade de válvulas e de hidrantes	*Decision Tree, Random Forest, AdaBoost, RUSBoost*
Kumar *et al*. (2018)	Diâmetro, Idade, Material, # de falhas, Pressão hidráulica, Densidade de quebras, Tráfego viário	*Decision Tree, Random Forest, AdaBoost, Logistic Regression Model*
Alizadeh *et al*. (2019)	Diâmetro, Idade, Material, Comprimento, Pressão hidráulica	*Artificial Neural Network, Support Vector Regression, Gaussian Regression*
Chen *et al*. (2019)	Tipo do solo, Tipo da superfície, Umidade do solo, Temperatura, Precipitação, Variáveis demográficas	*Decision Tree, Random Forest, Generalised linear models*

Continuação da Tabela 3.6 – Síntese da bibliografia pesquisada e estudada no contexto da presente pesquisa sobre falha em redes de água, sumarizando-se variáveis e métodos empregados.

Kerwin *et al*. (2019)	Diâmetro, Idade, # de falhas, Tipo do solo, Tráfego viário	*Artificial Neural Network*
Wols *et al*. (2019)	Diâmetro, Idade, Material, Comprimento, Umidade do solo, Temperatura, Precipitação	*Gradient-Boosted Tree, Random Forest, AdaBoost, Extra Trees Regression, Linear Regression*
Robles-Velasco *et al*. (2020)	Diâmetro, Idade, Material, Comprimento, # de falhas, Pressão hidráulica, Densidade de ligações	*Support Vector Machine, Logistic Regression Model*
Snider e McBean (2020)	Diâmetro, Idade, Comprimento, Umidade do solo, Proteção catódica do tubo	*XGBoost*
Almheiri *et al*. (2020)	Diâmetro, Idade, Material, Comprimento, # de falhas	*Artificial Neural Network, Ensemble Decision Tree, Ridge Regression*
Giraldo-González e Rodrígues (2020)	Diâmetro, Idade, Material, # de falhas, Uso da terra, Umidade do solo, Resistividade do solo, Precipitação, Quantidade de válvulas e de hidrantes	*Gradient-Boosted Tree, Bayes, Support Vector Machine, Artificial Neural Network, Linear, Poisson, Evolutionary Polynomial Regressions*
Snider e McBean (2020b)	Diâmetro, Material, Comprimento, # de falhas, Pressão hidráulica, Tipo do solo, Proteção catódica do tubo	*Gradient-Boosted Tree, Survival Analysis*

Considerando os recursos dispendidos e necessários para a fiabilidade da infraestrutura, é salutar aplicar ferramentas que permitam fortalecer os processos de gestão e tomada de decisão em prol da racionalização dos investimentos em sistemas de abastecimento de água. O uso de técnicas de aprendizado de máquina tem permeado abordagens recentes sobre predição de falha em redes. Neste sentido, constata-se que a existência de dados cadastrais e históricos operacionais e de manutenção em Sistemas de Informação Geográfica (SIG) são fundamentais para a aplicação dos modelos preditivos. Soma-se à base de dados as informações sobre a dinâmica operacional das redes, o que põe em evidência o valor agregado por meio de modelos hidráulicos calibrados (Robles-Velasco *et al*., 2020; Snider e McBean, 2020b), uma vez que informações sobre pressões operacionais mínimas, máximas, médias e sua amplitude, bem como vazões, velocidades e perdas de carga nas tubulações da rede são obtidas por meio da construção e calibração desses modelos (Itonaga, 2005).

É unanimidade entre os estudos constantes da revisão bibliográfica a relevância da aplicação de modelos preditores à falha em tubos da rede de distribuição de água que podem contribuir com gestão da infraestrutura em prol da economicidade e eficiência operacional. Em geral, tais estudos versam sobre rompimentos propriamente ditos ou vazamentos visíveis. Todavia, além dos rompimentos de tubulação de distribuição, os vazamentos em ramais/ligações de serviço de conexão ao hidrômetro do cliente representam parcela relevante das perdas reais em sistemas de abastecimento de água, principalmente os não visíveis. Tais vazamentos demandam esforços significativos tanto em termos de Controle Ativo de Vazamentos, quanto em reparos.

Embora vários estudos recentes tenham abordado a predição de falha nas redes de distribuição propriamente ditas, não se encontrou estudos específicos sobre vazamentos e rompimentos em ramais de serviço e de conexão entre a rede e o hidrômetro dos clientes. Portanto, esta pesquisa visa contribuir com pesquisas direcionadas à essa lacuna observada.

4. METODOLOGIA

Este capítulo descreve a metodologia e procedimentos realizados nesta pesquisa para conduzir ao cumprimento dos objetivos propostos. Primeiramente a área de estudo é caracterizada, seguida da apresentação da base de dados utilizada.

A metodologia utilizada na presente pesquisa baseia-se na aplicação das etapas propostas por Campesato (2020) quanto às ações para implementação de modelos de Aprendizado de Máquina:

- Obter o conjunto de dados;
- Realizar a limpeza da base de dados (*data cleaning* ou *data cleansing*). Este processo consiste em detectar, corrigir ou até remover registros corrompidos, incompletos, incorretos, duplicados, irrelevantes, ou imprecisos, de um banco de dados;
- Selecionar variáveis para aplicação dos modelos;
- Aplicar redução de dimensionalidade;
- Selecionar algoritmo;
- Selecionar dados de treinamento e dados de teste;
- Treinar o modelo;
- Testar o modelo;
- Ajustar o modelo; e,
- Avaliar resultados obtidos.

4.1. ÁREA DE ESTUDO E BASE DE DADOS

A área de estudo é o Distrito Federal (DF), sob prestação de serviços de abastecimento de água e de esgotamento sanitário da Caesb, Companhia de Saneamento ambiental do Distrito Federal. Para aplicação na presente pesquisa, serão utilizados dados disponíveis sobre as instalações físicas das tubulações de redes de distribuição de água, tais como diâmetro, extensão, material e data de implantação, somados às informações cadastrais e comerciais das ligações de água.

Além de dados contidos em Sistemas de Informações Geográficas (SIG), a presente pesquisa utilizará modelos hidráulicos construídos e calibrados entre 2017 e 2018 de 5,7

mil km de redes de distribuição de água operados pela Caesb, o que representa aproximadamente 60% da rede em operação. Essa área de cobertura abrange 540 mil ligações de água, que possuem registros de ocorrências de vazamentos visíveis reparados em seus ramais/ ligações de serviço entre a rede de distribuição de água e o hidrômetro do cliente.

A base de dados sobre vazamentos visíveis reparados contempla o período entre janeiro de 2015 e abril de 2020. Os vazamentos foram georreferenciados e indexados por ligação de água em SIG. Além desses dados, os modelos hidráulicos das redes de distribuição (em *Epanet* e *InfoWater*) presta a inserção das seguintes grandezas hidráulicas: pressões operacionais; velocidades; vazões; e, perda de carga na tubulação que provê derivação ao ramal. Portanto cada ramal de ligação possui a informação se houve reparo de vazamento visível ou não no período, além das características sobre rede de distribuição que dá origem a cada ramal.

A Figura 4.1, a seguir, apresenta a localização de Brasília/DF no Brasil, bem como identifica as redes modeladas utilizadas e ilustra a distribuição de um dos parâmetros utilizados, as pressões máximas operacionais observadas nas redes de distribuição.

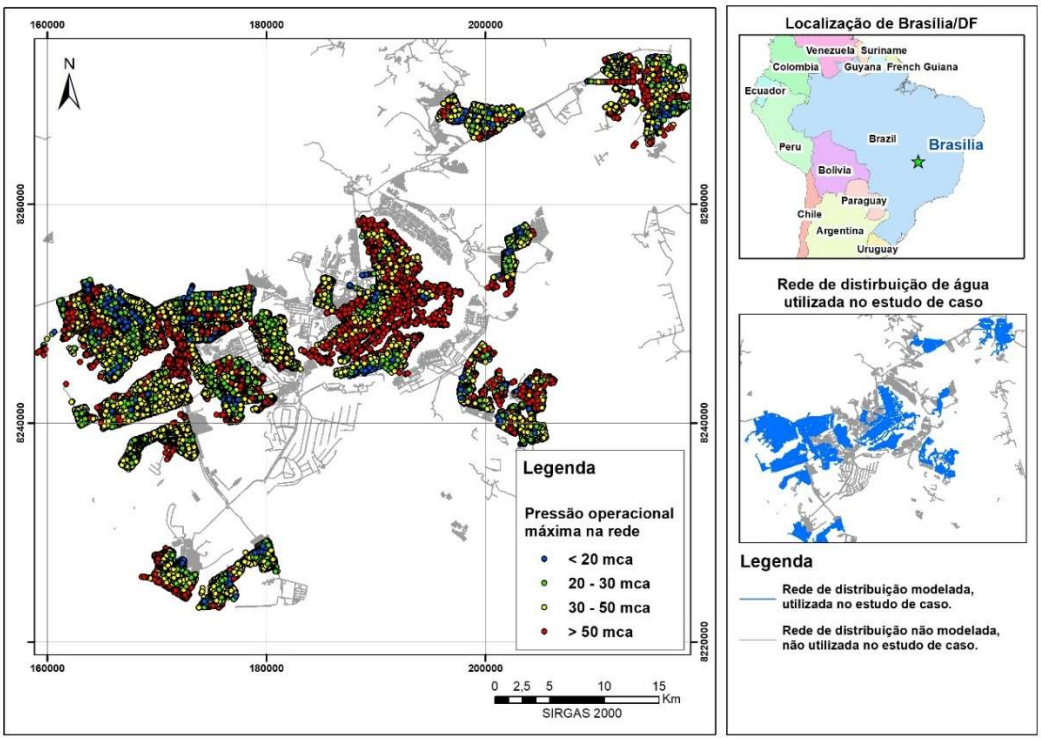


Figura 4.1 – Brasília/DF, identificação de áreas com modelos hidráulicos das redes de distribuição e exemplo de dados obtidos por meio da simulação (pressões máximas operacionais).

A Base de Dados é composta por 38 variáveis, agrupadas em quatro grupos de acordo com a origem da variável, aspectos operacionais (10 variáveis), aspectos físicos (07), aspectos comerciais (04), e aspectos ambientais (17). Tal agrupamento foi inspirado por meio da classificação das variáveis amplamente aplicada nas publicações constantes da Revisão Bibliográfica. Com exceção da variável "RAP", que identifica qual o reservatório responsável pelo abastecimento da ligação, em formato de texto, todas as demais variáveis são do tipo numérica, adequadas ao processamento em qualquer modelo de aprendizado de máquina. A variável "LEAK" é a variável-alvo categórica (variável de saída) utilizada nos modelos, correspondendo a: [0] *ramais que não apresentaram vazamento visível reparado,* e, [1] *ramais que apresentaram vazamento visível reparado*. A relação de variáveis é apresentada na

Tabela 4.1. Quatro das seis variáveis de ordem operacional são provenientes de simulações hidráulicas das redes de distribuição. Tabela 4.2, por sua vez, apresenta a relação fonte do dado x variável.

A variável "RAP" é utilizada para possibilitar a distinção entre dados provenientes de 14 áreas de abastecimento por reservatório apoiado no DF. Foram utilizados os seguintes reservatórios: RAP.BRT.001, RAP.CEI.001, RAP.LSL.002, RAP.MNT.001, RAP.MNT.002, RAP.PPL.001, RAP.PPL.002, RAP.PRN.002, RAP.RCE.001, RAP.SAM.001, RAP.SAM.002, RAP.SSB.001, RAP.SSB.002, e RAP.VCP.001. Os modelos foram processados por área de atendimento de reservatório para fins de avaliação da performance por região, considerando eventuais distinções entre características dessas localidades que não estejam estruturadas para inserção como uma variável para processamento.

Tabela 4.1– Variáveis utilizadas nos modelos de Aprendizado de Máquina.

Variáveis constantes da Base de Dados (*Variáveis obtidas por simulações hidráulicas)			
Origem da variável	Variável	Descrição	Tipo da variável
Aspectos operacionais	LEAK	Identifica se a ligação teve vazamento reparado	Núm. discreta [0 ou 1]
	*MOD_AVG_HLOSS	Perda de carga média no tubo (m/km) que provê derivação à ligação	Núm. contínua
	*MOD_AVG_PRESS	Pressão média no tubo (mca) que provê derivação à ligação	Núm. contínua
	*MOD_AVG_VELOC	Velocidade média no tubo (m/s) que provê derivação à ligação	Núm. contínua
	*MOD_MAX_PRESS	Pressão máxima no tubo (mca) que provê derivação à ligação	Núm. contínua
	*MOD_MIN_PRESS	Pressão mínima no tubo (mca) que provê derivação à ligação	Núm. contínua
	*MOD_RAN_PRESS	Variação de pressão no tubo (mca) que provê derivação à ligação	Núm. contínua
	VALVE	Ligação atendida por VRP	Núm. discreta [0 ou 1]
	BOOSTER	Ligação atendida por EBO	Núm. discreta [0 ou 1]
	RAP	Reservatório Apoiado responsável pelo atendimento da ligação	Categórica, texto
Aspectos físicos	CONN_AGE	Idade da conexão/ramal de ligação (ano)	Núm. discreta
	MAT_CI_FF	O material do tubo que provê derivação à ligação é Ferro Fundido	Núm. discreta [0 ou 1]
	MAT_HDPE_PEAD	O material do tubo que provê derivação à ligação é PEAD	Núm. discreta [0 ou 1]
	MAT_MPVC_DEFOFO	O material do tubo que provê derivação à ligação é DEFOFO	Núm. discreta [0 ou 1]
	MAT_PVC	O material do tubo que provê derivação à ligação é PVC	Núm. discreta [0 ou 1]
	WN_AGE	Idade do tubo (ano) que provê derivação à ligação	Núm. discreta
	WN_DIAMETE	Diâmetro do tubo (mm) que provê derivação à ligação	Núm. discreta
Aspectos comerciais	USE_COM	Cliente com uso comercial	Núm. discreta [0 ou 1]
	USE_IND	Cliente com uso industrial	Núm. discreta [0 ou 1]
	USE_PUB	Cliente com uso público	Núm. discreta [0 ou 1]
	USE_RES	Cliente com uso residencial	Núm. discreta [0 ou 1]
Aspectos ambientais	PAVING_ASPHALT	Ligação com asfalto à porta	Núm. discreta [0 ou 1]
	PAVING_NO_PAVING	Ligação sem asfalto à porta	Núm. discreta [0 ou 1]
	ROUTE_BUS	Ligação em frente a via de circulação de ônibus	Núm. discreta [0 ou 1]
	ROUTE_TYPE_ART	Ligação em frente a via arterial	Núm. discreta [0 ou 1]
	ROUTE_TYPE_COLEC	Ligação em frente a via coletora	Núm. discreta [0 ou 1]
	ROUTE_TYPE_FAST	Ligação em frente a via rápida	Núm. discreta [0 ou 1]
	ROUTE_TYPE_HIGH	Ligação em frente a via expressa/rodovia	Núm. discreta [0 ou 1]
	ROUTE_TYPE_LOCAL	Ligação em frente a via local	Núm. discreta [0 ou 1]
	ROUTE_VELOCITY	Velocidade máxima de trânsito na via (km/h)	Núm. discreta [0 ou 1]
	SLOPE	Declividade do terreno (%) sob o tubo que provê derivação à ligação	Núm. contínua
	SOIL_CX	Ligação sob cambissolo háplico	Núm. discreta [0 ou 1]
	SOIL_FF	Ligação sob plintossolo pétrico	Núm. discreta [0 ou 1]
	SOIL_GX	Ligação sob gleissolo háplico	Núm. discreta [0 ou 1]
	SOIL_LV	Ligação sob latossolo vermelho	Núm. discreta [0 ou 1]
	SOIL_LVA	Ligação sob latossolo vermelho amarelo	Núm. discreta [0 ou 1]
	SOIL_NV	Ligação sob nitossolo vermelho	Núm. discreta [0 ou 1]
	SOIL_RQ	Ligação sob neossolo quartzarênico	Núm. discreta [0 ou 1]

Tabela 4.2 – Fonte dos dados que originaram as variáveis aplicadas nos modelos de Aprendizado de Máquina.

Fonte do dado	Variável
Cadastro técnico e operacional da Companhia	LEAK
	VALVE
	BOOSTER
	RAP
	CONN_AGE
	MAT_CI_FF
	MAT_HDPE_PEAD
	MAT_MPVC_DEFOFO
	MAT_PVC
	WN_AGE
	WN_DIAMETE
	USE_COM
	USE_IND
	USE_PUB
	USE_RES
Cadastro técnico e operacional da Companhia Simulação hidráulica das redes de distribuição	MOD_AVG_HLOSS
	MOD_AVG_PRESS
	MOD_AVG_VELOC
	MOD_MAX_PRESS
	MOD_MIN_PRESS
	MOD_RAN_PRESS
Processamento de Modelo Digital do Terreno a partir de curvas de nível disponíveis para o DF em Escala 1:10.000	SLOPE
Núcleo de Geoprocessamento da Companhia Urbanizadora da Nova Capital do Brasil (Novacap)	PAVING_ASPHALT
	PAVING_NO_PAVING
	ROUTE_BUS
	ROUTE_TYPE_ART
	ROUTE_TYPE_COLEC
	ROUTE_TYPE_FAST
	ROUTE_TYPE_HIGH
	ROUTE_TYPE_LOCAL
	ROUTE_VELOCITY
Levantamento de reconhecimento dos solos do Distrito Federal, Empresa Brasileira de Pesquisa Agropecuária (Embrapa)	SOIL_CX
	SOIL_FF
	SOIL_GX
	SOIL_LV
	SOIL_LVA
	SOIL_NV
	SOIL_RQ

4.2. PREPARAÇÃO DE DADOS, ANÁLISE EXPLORATÓRIA DE DADOS E REDUÇÃO DE DIMENSIONALIDADE

A compilação dos dados foi realizada em SIG, integrando-se dados por meio do *software* ArcMap 10.6.1. Outras fases: pré-processamento, Análise Exploratória de Dados, aplicação (treinamento, teste, Validação Cruzada) de modelos de Aprendizado de Máquina, e Análise de Desempenho ocorreu em ambiente Anaconda utilizando-se Python 3.7 e suas bibliotecas, tais como *Numpy*, *Pandas*, *Matplotlib*, *Seaborn*, *SKLearn*, e *Imblearn*.

O Banco de Dados foi compilado após processamento para remoção de dados inconsistentes, preenchimento de falhas, e remoção de *outliers* por meio da aplicação do método dos Quartis (Tukey, 1977). Em seguida, a Base de Dados foi submetida a processos de Análise Exploratória dos Dados para compreensão das características básicas dos dados disponíveis. Concluindo esta etapa, aplicou-se Redução de Dimensionalidade à Base de Dados para avaliação da distribuição da variância dos dados considerando a variável-alvo: [0] *ramais que* ***não*** *apresentaram vazamento visível reparado*, e, [1] *ramais que apresentaram vazamento visível reparado*.

4.3. APLICAÇÃO DE MODELOS PREDITORES POR APRENDIZADO DE MÁQUINA

O presente estudo aplicou os seguintes métodos de Aprendizado de Máquina por classificação à Base de Dados: *Linear Svm*, *Radial Svm*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, e, *Random Forest*. Em seguida, *Ensamble Learning Models* também foram aplicados: *Bagged K-nearest Neighbors*, *Bagged Decision Tree*, *Adaboost*, *Gradient Boosting* e *XGBoost*. A classificação foi dada por meio de atributo (LEAK) que diferenciou os ramais que apresentaram vazamento visível reparado no período de coleta de dados, variável-alvo: [0] *ramais que* ***não*** *apresentaram vazamento visível reparado*, e, [1] *ramais que apresentaram vazamento visível reparado*. O treinamento de todos os modelos foi baseado em Validação Cruzada, utilizando-se 10 *folders*.

A utilização de diferentes técnicas de Aprendizado de Máquina se justifica sob a ótica da avaliação da performance de diferentes métodos e que, havendo possibilidade de

apuração computacional de distintos métodos, é salutar seus usos e comparações em busca da obtenção das melhores predições, que repercutem diretamente em maior acurácia para o processo de tomada de decisão.

O modelo que apresentou melhor performance foi processado por uma última etapa de aperfeiçoamento, expondo-o a hiper-parametrização para eventuais ganhos finais de desempenho.

4.4. AVALIAÇÃO DE DESEMPENHO DOS MODELOS PREDITORES POR APRENDIZADO DE MÁQUINA

Todos os modelos aplicados foram avaliados por meio das seguintes métricas, vide Item 3.5:

- Acurácia *(Score)* final;
- Matriz de Confusão; e,
- Importância das variáveis.

4.5. AVALIAÇÃO DE FATORES INTERVENIENTES À FALHA E POSSIBILIDADES DE APLICAÇÃO DOS MODELOS NOS PROCESSOS DE GESTÃO DE PERDAS DE ÁGUA

A aplicação dos modelos permite o ordenamento dos fatores intervenientes à falha em ordem de importância à classificação de vazamento. Portanto, foi avaliado e discutido quais fatores são indutores do desenvolvimento de vazamentos nos ramais estudados constantes da Base de Dados disponível, o que pode fornecer informações úteis ao processo de gestão da infraestrutura.

Por fim, é discutido os potenciais de aplicação de modelos de Aprendizado de Máquina em processos de Gestão de Perdas de Água, no que tange a identificação de ramais mais vulneráveis ao desenvolvimento de vazamentos não visíveis que possam ser investigados por meio de atividades de Pesquisa de Vazamentos.

A classificação por Aprendizado de Máquina dos ramais em [0] *ramais que* ***não*** *apresentaram vazamento visível reparado*; e, [1] *ramais que apresentaram vazamento visível reparado*; permitirá, por meio da Matriz de Confusão, que os ramais sejam

agrupados em quatro conjuntos: Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo, o que permite aprofundamento sobre a performance dos modelos e avaliação sobre potenciais aplicações.

4.6. FLUXOGRAMA DA PESQUISA

A Figura 4.2, a seguir apresenta o Fluxograma da proposta de pesquisa.

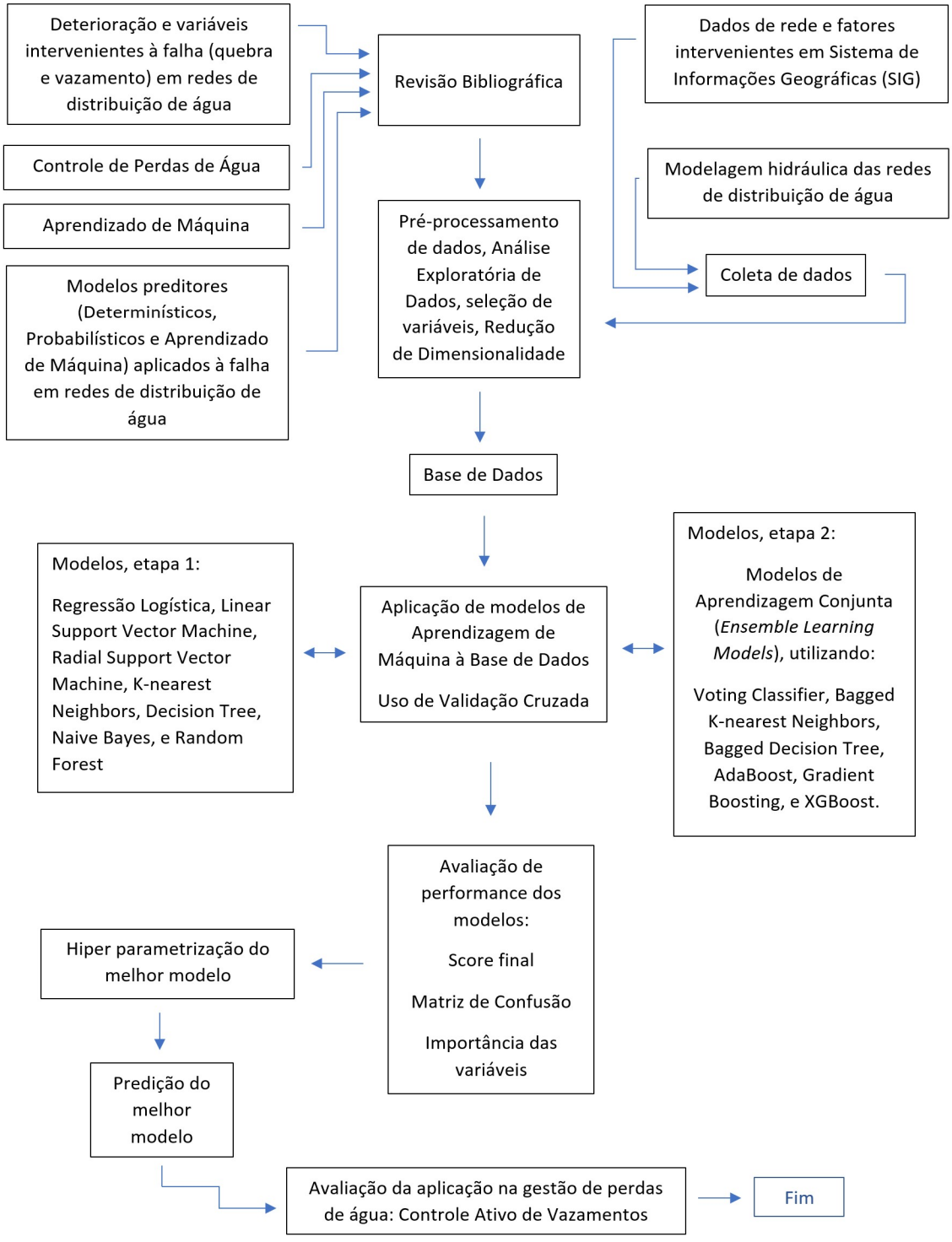


Figura 4.2 – Fluxograma da proposta de pesquisa.

5. RESULTADOS E DISCUSSÕES

Este capítulo apresenta os principais resultados obtidos por meio deste trabalho, bem como as discussões pertinentes a cada tópico, aplicada a metodologia descrita no Capítulo 4.

5.1. ANÁLISE EXPLORATÓRIA DOS DADOS

Após a obtenção do conjunto de dados em ambiente SIG, segundo Campesato (2020), é necessário realizar a limpeza da base de dados (*data cleaning* ou *data cleansing*). Neste contexto, duas ações principais foram adotadas para compor a base de dados. A primeira medida foi a remoção de registros de ligações com mais de um vazamento realizado no período de análise, considerando que a recorrência pode ser desdobramento da qualidade/ método de reparo realizado, o que não é objeto da presente pesquisa uma vez que não há dados que possam sustentar o aprofundamento neste ponto.

A segunda medida é a remoção de *outliers* da base de dados. *Outliers* são registros de dados que diferem significativamente de todos os outros, se distinguindo em uma ou mais características. Em outras palavras, um *outlier* é um valor que foge da normalidade e pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas analíticos. Há estudos em que a identificação de *outliers* é buscada, por exemplo, análises de fraude. Em outros, como no caso desta investigação, tais registros geram ruído à aplicação dos modelos de aprendizado de máquina. Portanto, os *outliers* foram removidos utilizando-se o método dos quartis.

A remoção de *outliers* foi aplicada nas variáveis numéricas contínuas, com ação principalmente nas seguintes variáveis: MOD_AVG_PRESS, MOD_MIN_PRESS, MOD_MAX_PRESS, MOD_RAN_PRESS, SLOPE e DIAMETE. As variáveis provenientes dos modelos hidráulicos, com pressões observadas acima de 100 mca foram removidas. Tais condições não são observadas na rede e distribuição propriamente dita no DF, dado a instrumentação para controle de pressão existente. Uma análise sobre as

ligações que obtiveram pressões elevadas mostrou que forma casos em que a ligação se associou indevidamente a adutoras de alta pressão, à montante de VRPs, ou a linhas de recalque no sistema. Não constitui escopo da presente pesquisa refinar o georreferenciamento de ligações que possa ser aprimorado. Portanto, tal condição foi corrigida pela remoção dos *outliers*. O mesmo procedimento aborda a presença de ligações provenientes de redes de grandes diâmetros. Ligações deste tipo são raras, dado que tais diâmetros normalmente constituem redes primárias, mas podem ocorrer devido fatores diversos. A remoção de outliers retirou ligações derivadas de redes de grandes diâmetros, considerando, também, a questão já citada sobre o posicionamento da ligação.

Embora a remoção de *outliers* tenha sugerido, matematicamente, que a declividade deveria estar compreendida entre 0 e 5%, sabidamente há regiões no DF com habitações em condições de declividade superior. Portanto, para o caso da declividade ("SLOPE"), não se manteve a exclusão das ligações em áreas de elevada declividade. A Base de Dados utilizada nos modelos manteve esses registros. As variáveis sobre a idade da tubulação da rede de distribuição que presta origem à ligação de água ("WN_AGE") e em relação a idade da ligação de água ("CONN_AGE") não foram afetadas pela remoção de *outliers*, dado que a consistência de datas de implantação no banco de dados foi previamente avaliada, além de ser objeto de análise de consistência por parte do banco de dados corporativo da Concessionária.

A Figura 5.1 – Histogramas de variáveis antes da aplicação da remoção de outliers (em cinza), após a remoção de outliers (em azul). apresenta os histogramas de variáveis, antes e após a remoção de *outliers*. As Figuras Figura 5.2 e Figura 5.3 apresentam os histogramas em conjunto com o diagrama de caixa para as variáveis antes da remoção de *outliers*; as Figuras Figura 5.4 e Figura 5.5, por sua vez, apresentam os histogramas em conjunto com o diagrama de caixa para as variáveis após a remoção de *outliers*. A alta frequência de tubulações na faixa de diâmetro de 60mm sugere o uso preferencial de tubos de PVC em 60mm e de PEAD em 63mm para as redes distribuição secundária. Pressões médias e máximas giram em torno de 30mca; quanto às mínimas, a maior frequência é 25mca. O range de pressões mais frequente é inferior a 15mca, as ligações de água predominam idade inferior a 30 anos. Percebe-se que a frequência das tubulações entre

40 e 50 anos é superior a frequência da idade das ligações para o mesmo período, o que indica que a renovação das ligações é superior a renovação da rede, mas também pode indicar que muitas áreas foram efetivamente ocupadas e passaram por adensamento após a instalação das redes; portanto, tal comportamento é normal.

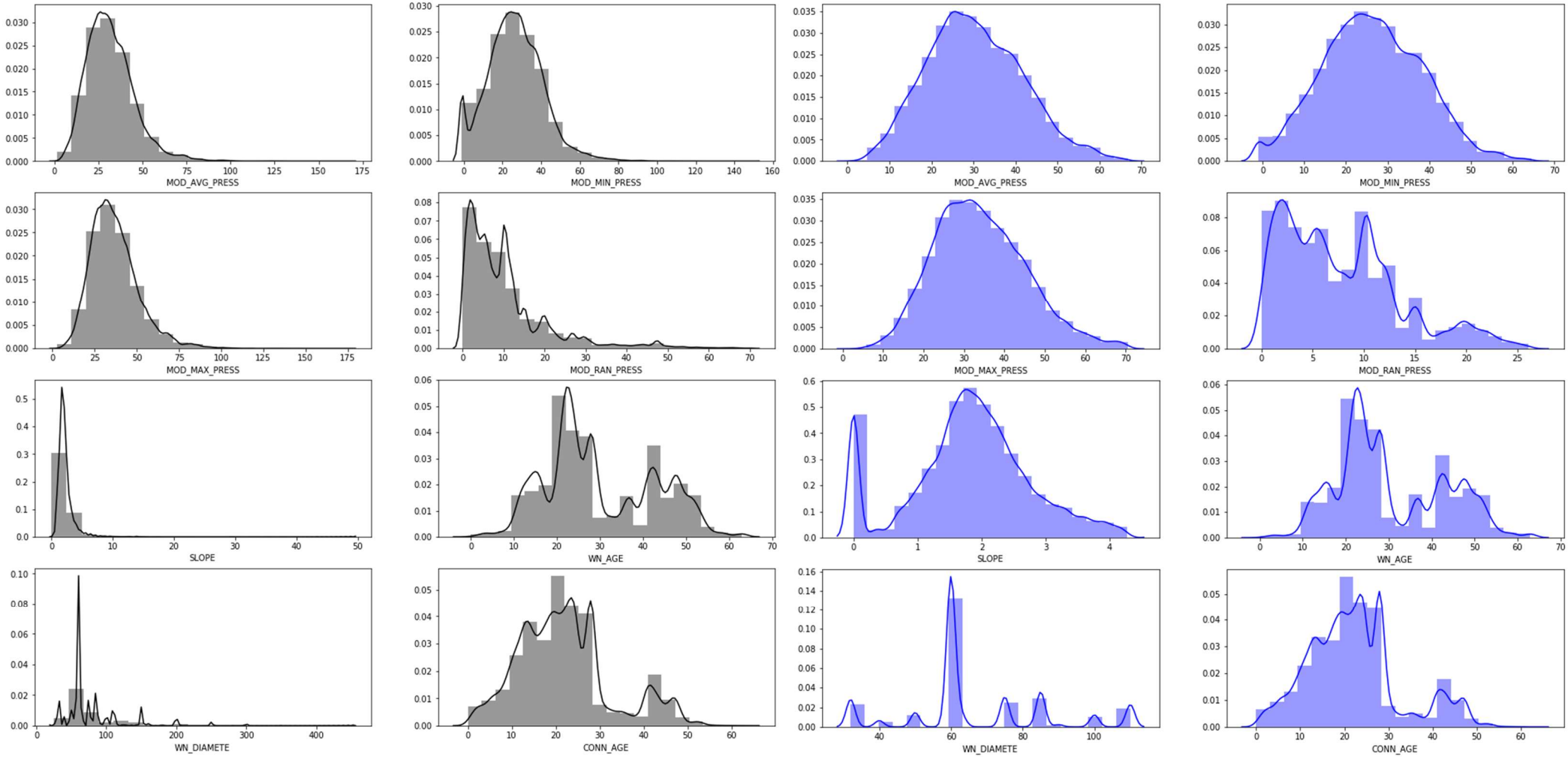


Figura 5.1 – Histogramas de variáveis antes da aplicação da remoção de outliers (em cinza), após a remoção de outliers (em azul).

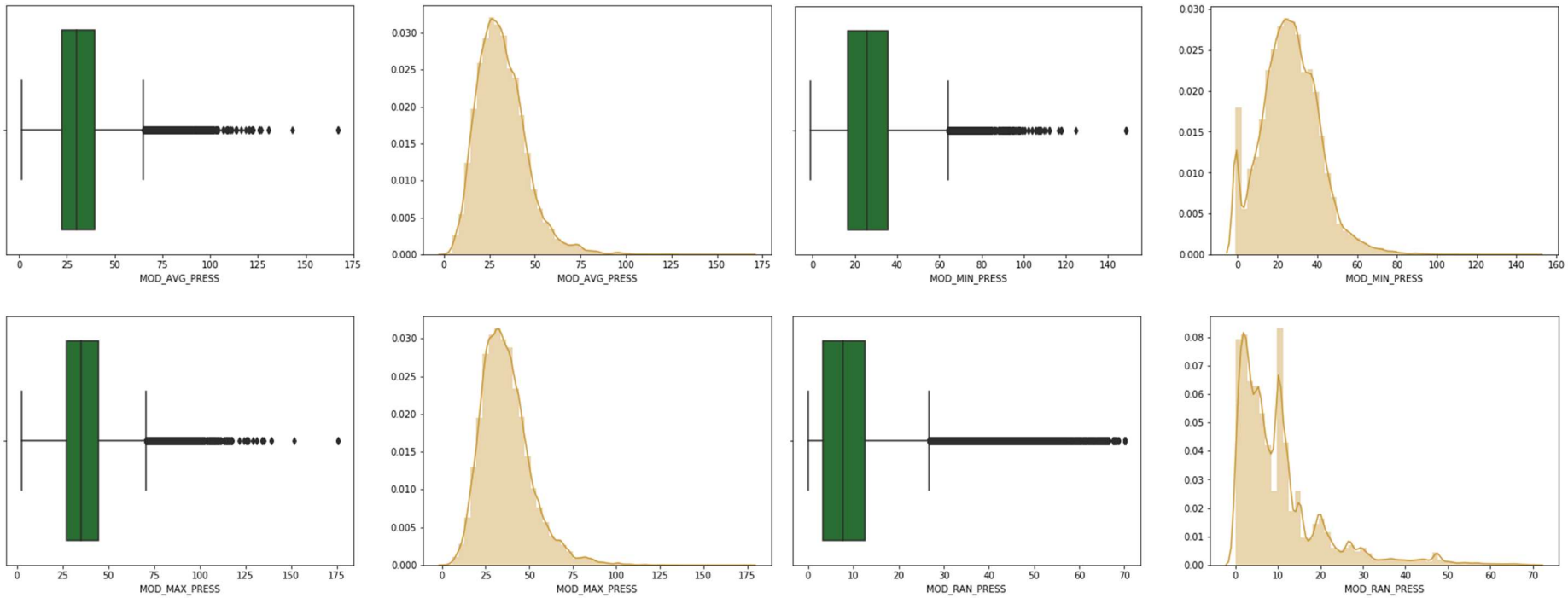


Figura 5.2 – Histogramas e diagrama de caixa das variáveis antes da remoção de *outliers*.

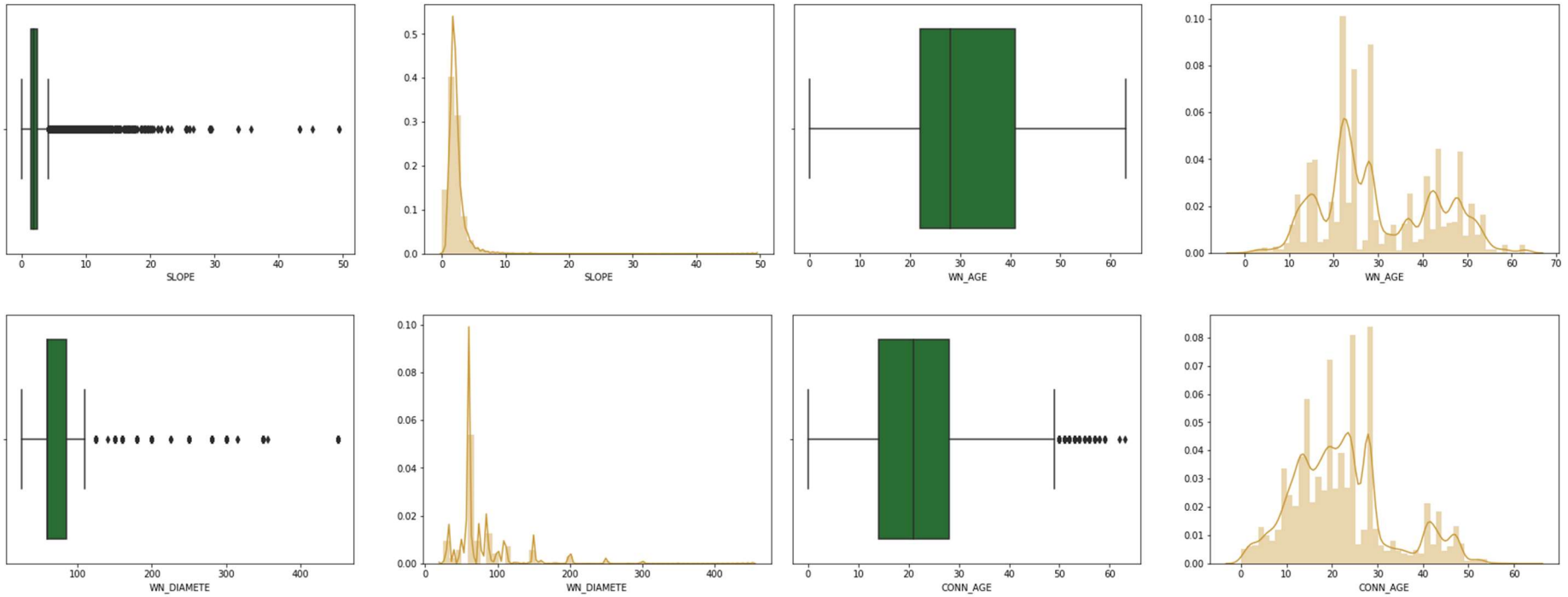


Figura 5.3 – Histogramas e diagrama de caixa das variáveis antes da remoção de *outliers*.

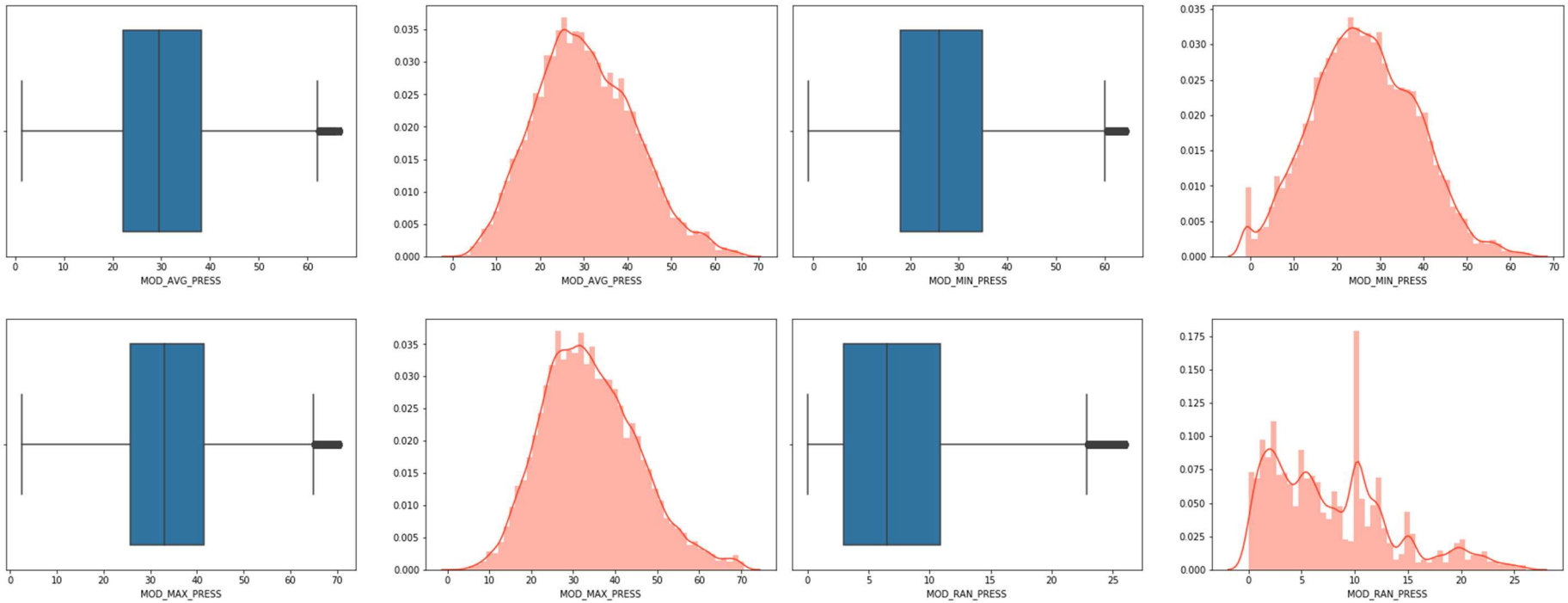


Figura 5.4 – Histogramas e diagrama de caixa das variáveis após a remoção de *outliers*.

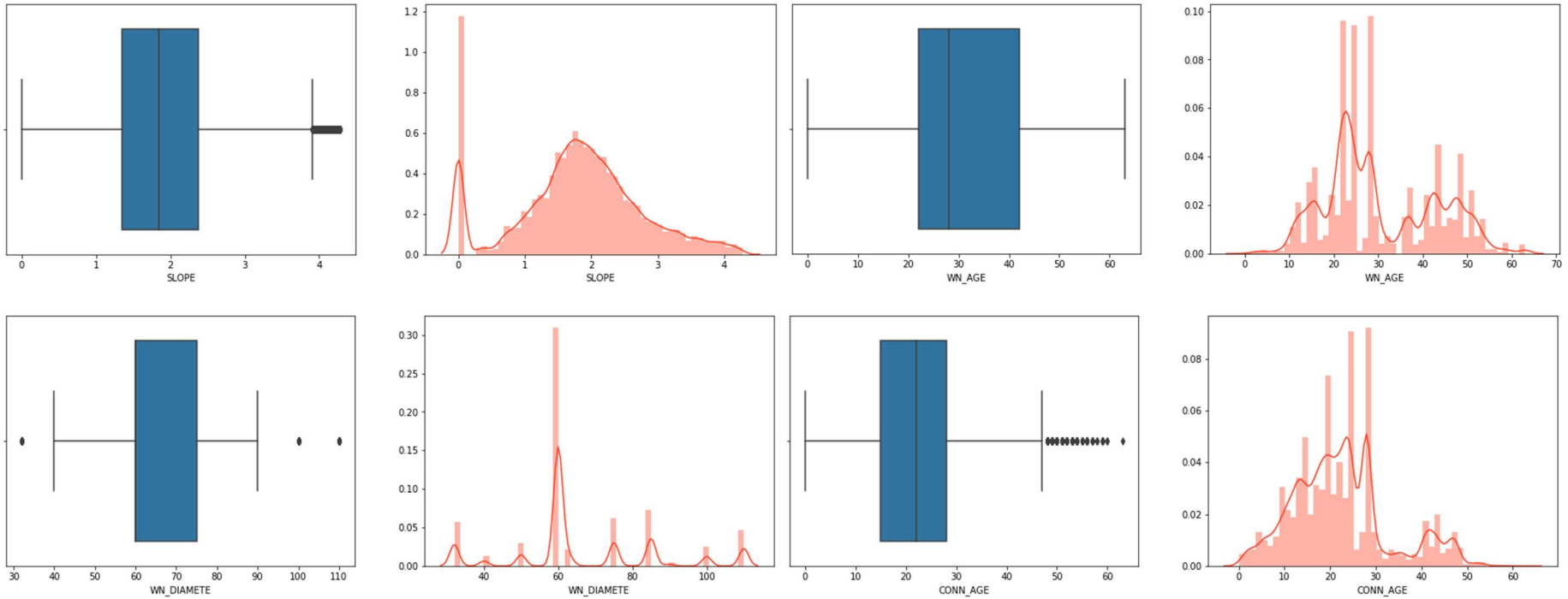


Figura 5.5 – Histogramas e diagrama de caixa das variáveis após a remoção de *outliers*.

A Base de Dados consistida e aplicada possui 348.736 registros, dos quais 66.917 representam ligações em que houve vazamento visível reparado no período de análise; assim, 281.819 índices são de ligações de água sem ocorrência para vazamento. Portanto, a base de dados não é balanceada, sendo a classe de registros com vazamento visível reparado igual a 19,19% do total. Segundo Kaur *et al*. (2019), uma forma de tratar bancos de dados não balanceados é por meio da seleção de amostra da classe com maior presença, buscando-se compor um banco de dados em que cada classe possua a mesma representatividade. Em que pese o balanceamento não ser uma obrigatoriedade, optou-se por realizar as análises com o conjunto de dados balanceado. Assim, as classes [0] *ramais que* ***não*** *apresentaram vazamento visível reparado*, e, [1] *ramais que apresentaram vazamento visível reparado*, foram balanceadas com a seleção integral dos registros da base de dados [1], vazamentos, e os registros [0] foram selecionados aleatoriamente, na mesma quantidade.

A única variável categórica do tipo texto é "RAP", que identifica qual reservatório apoiado abastece a ligação. Essa variável não é utilizada como entrada para processamento nos modelos de Aprendizado de Máquina. Esta variável é utilizada para que a Base de Dados possa ser processada por reservatório. As áreas de abastecimento de 14 reservatórios no DF foram consideradas: RAP.BRT.001, RAP.CEI.001, RAP.LSL.002, RAP.MNT.001, RAP.MNT.002, RAP.PPL.001, RAP.PPL.002, RAP.PRN.002, RAP.RCE.001, RAP.SAM.001, RAP.SAM.002, RAP.SSB.001, RAP.SSB.002, e RAP.VCP.001. A presença de ramais que apresentaram vazamento reparado oscila entre 14 e 31% nas áreas desses reservatórios, conforme apresentado na

Tabela 5.1.

Tabela 5.1 – Percentual de ramais com e sem vazamento visível reparado nas áreas de atendimento dos reservatórios utilizados nesta investigação.

RAP	Ramais que **não** apresentaram vazamento visível reparado, LEAK = 0	Ramais que apresentaram vazamento visível reparado, LEAK = 1
RAP.BRT.001	79,00%	21,00%
RAP.CEI.001	84,72%	15,28%
RAP.LSL.002	72,88%	27,12%
RAP.MNT.001	80,42%	19,58%
RAP.MNT.002	85,51%	14,49%
RAP.PPL.001	76,61%	23,39%
RAP.PPL.002	76,39%	23,61%
RAP.PRN.002	79,37%	20,63%
RAP.RCE.001	80,37%	19,63%
RAP.SAM.001	80,23%	19,77%
RAP.SAM.002	77,21%	22,79%
RAP.SSB.001	68,53%	31,47%
RAP.SSB.002	70,99%	29,01%
RAP.VCP.001	83,88%	16,12%
Total	80,81%	19,19%

Quanto às variáveis numéricas discretas, observa-se a presença significativa (37,66%) de ligações atendidas após redução de pressão por meio de Válvulas Redutoras de Pressão (VRP), 22,51% das ligações são abastecidas por meio do uso de elevatória tipo *Booster* na rede de distribuição e, então, 39,83% das ligações possuem a pressão de serviço regida por meio de Reservatório Apoiado (RAP). Os 60% das ligações atendidas por meio de equipamentos que visam, entre outros aspectos, reduzir as pressões operacionais das redes de distribuição de água é uma boa prática para controle de perdas. 88% das ligações possuem asfalto à porta, o principal tipo de via à porta dos clientes são as vias locais (74%), o latossolo vermelho é o principal tipo de solo encontrado na região (80%), o principal tipo de uso dos imóveis é residencial (88%), e o principal material utilizado nas redes de distribuição é o PVC (57%). A

Tabela 5.2, a seguir, apresenta o percentual de ligações agrupadas a cada variável categórica.

Tabela 5.2 – Percentual de ligações por variável categórica.

Variáveis categóricas	**Percentual de ligações/ variável (%)**
Ligação atendida por EBO	22,51
Ligação atendida por VRP	37,66
Ligação com asfalto à porta	87,91
Ligação sem asfalto à porta	12,09
Ligação em frente a via expressa/rodovia	0,36
Ligação em frente a via arterial	5,67
Ligação em frente a via coletora	19,07
Ligação em frente a via rápida	0,44
Ligação em frente a via local	74,46
Ligação em frente a via de circulação de ônibus	7,75
Ligação sob cambissolo háplico	6,17
Ligação sob plintossolo pétrico	0,00
Ligação sob gleissolo háplico	1,48
Ligação sob latossolo vermelho	80,46
Ligação sob latossolo vermelho amarelo	10,63
Ligação sob nitossolo vermelho	0,19
Ligação sob neossolo quartzarênico	1,08
Ligação para uso comercial	11,24
Ligação para uso industrial	0,28
Ligação para uso público	0,51
Ligação para uso residencial	87,97
Ligação proveniente de rede em PVC	57,26
Ligação proveniente de rede em PEAD	16,69
Ligação proveniente de rede em DEFOFO	0,65
Ligação proveniente de rede em FF	25,29

Em continuidade à Análise Exploratória de Dados, a Base de Dados foi processada para compor uma Matriz de Correlação. Uma Matriz de Correlação é uma tabela que exibe os coeficientes de correlação, par a par, para as variáveis constantes de um banco de dados.

É uma ferramenta de síntese para resumir um conjunto de dados e permitir a visualização de padrões nos dados fornecidos. Cada célula em uma Matriz de Correlação contém o coeficiente de correlação entre um par de variáveis. A Figura 5.6 retrata a Matriz de Correlação para as 36 variáveis do Banco de Dados, com exceção da variável-alvo e da variável categórica.

Ligações atendidas por meio de pressurização direta da rede, por meio de *Booster*, possuem correlação negativa quanto às pressões máximas, mínimas e médias. Tal condição é reflexo das características de operação de tais equipamentos, em que a modulação da rotação das bombas também repercute nas pressões da rede, que tendem a ter menores oscilações. As ligações atendidas por Válvulas Redutoras de Pressão, por sua vez, possuem correção negativa menos acentuada àquelas atendidas por *Booster*. Esta observação se dá em decorrência das VRPs não possuírem a ampla gama de pressões de jusante proporcionadas pelos EBOs. É normal as VRPs não modularem pressão de jusante; aquelas que o fazem, normalmente modulam apenas duas vezes ao dia. Assim, a Base de Dados mostra que as ligações atendidas por meio de VRPs e EBOs possuem menores pressões máximas, mínimas e médias. Em complemento a esta análise, a Matriz de Correlação também permite identificar que a presença de EBOs e VRPs na rede de distribuição possui correlação negativa. Este comportamento é esperado dado que o design das redes de distribuição deve ser concebido para não demandar estes equipamentos em cascata por questões de eficiência energética.

As variáveis hidráulicas sobre pressões máximas, mínimas e média, possuem alta correlação entre si, o que é esperado em termos de funcionamento de redes de distribuição de água. Observa-se que áreas atendidas por redes com maior declividade possuem correlação positiva com essas variáveis, indicando que à medida que a declividade aumenta, as redes de distribuição são mais suscetíveis a operar com maiores pressões. Avaliando-se o comportamento das variáveis afetas à pavimentação das vias, compreende-se que ligações em frente às vias não pavimentadas e ligações em frente às vias pavimentadas possuem comportamento dicotômico. As ligações em frente às vias pavimentadas são mais antigas, possuem tubulações em operação a mais tempo, são cobertas por latossolo, possuem maiores diâmetros e pressões médias. As ligações

contíguas às vias não pavimentadas indicam que estas áreas receberam ligações de água recentemente, são provenientes de redes com maiores perdas de carga, maiores velocidades e maior variação entre pressões máximas e mínimas. Tal condição indica que essas ligações estão localizadas em áreas em processo de urbanização (a receber pavimentação) e que a rede de distribuição de água tende a ser instalada em diâmetros mínimos, além das pressões operacionais serem menores, o que indica que são áreas que pode haver, inclusive, intermitência no abastecimento.

Os tubos em ferro fundido possuem correlação positiva com a idade dos tubos, indicando que este material foi prioritariamente utilizado nos primórdios da implantação da infraestrutura. Percebe-se que à medida que materiais plásticos passaram a ser utilizados, a presença do ferro fundido foi reduzida, condição observada por meio da correlação negativa entre a presença desses materiais (PVC e PEAD) quanto à idade do tubo. Ligações antigas possuem correlação positiva com cobertura por latossolo, baixa declividade e tubos em ferro fundido. Ligações mais recentes possuem afinidade com tubos em PVC e PEAD, além de maiores declividades em cambissolo háplico.

Outra ferramenta eficaz para avaliar padrões e relacionamentos entre variáveis é o gráfico de pares (*pairs plot* ou *scatterplot matrix*), que permite a visualização da distribuição de variáveis únicas e, também, os relacionamentos entre duas variáveis, inclusive permite a distinção e apresentação de dois ou mais subtipos em uma variável-alvo.

A Figura 5.7 apresenta o gráfico de pares para as variáveis de pressões médias, máximas, mínimas e range de pressão. A Figura 5.8 ilustra as variáveis de idade da rede, diâmetro da rede, idade da ligação de água, e declividade, em formato de gráfico de pares. As Figuras Figura 5.7 e Figura 5.8 apresentam dados identificados por meio das classes [0] *ramais que* ***não*** *apresentaram vazamento visível reparado*, e, [1] *ramais que apresentaram vazamento visível reparado*, constantes da variável LEAK. Identifica-se que as ligações que apresentaram vazamento se sobressaem em pressões mínimas, máximas e médias a partir de 30 mca; abaixo deste patamar, as ligações que não apresentaram vazamento possuem maior densidade. Tal comportamento é esperado, dado

que maiores pressões de serviço naturalmente podem induzir a maior incidência de falha nos materiais. Conexões mais recentes, com menos de 10 anos, apresentam menor densidade para vazamentos. Por fim, declividade (até 5%), variação da pressão de serviço (range), e diâmetro apresentam baixa variação entre as classes.

As Figuras Figura 5.9, Figura 5.10, Figura 5.11, e Figura 5.12, contém, respectivamente, dados sobre os materiais PVC, PEAD, DEFOFO, e FF, confrontados, a cada figura, quanto ao desempenho dos demais materiais quanto às pressões operacionais. O PVC possui menor volume de registros para vazamento em pressões altas pressões, comparativamente aos demais materiais, embora apresente maior densidade na faixa de 40 mca. O PEAD é visivelmente mais sensível às pressões operacionais mais altas, comparativamente os demais materiais, assim como o DEFOFO, que adicionalmente imprime maior fragilidade às ligações em situação de maior variabilidade da pressão de atendimento da rede. As ligações atendidas a partir de derivação em FF, por sua vez, apresentaram menor susceptibilidade às pressões, mas fragilidade superior aos demais materiais quando a oscilação de pressões (range) é da ordem de 20 mca.

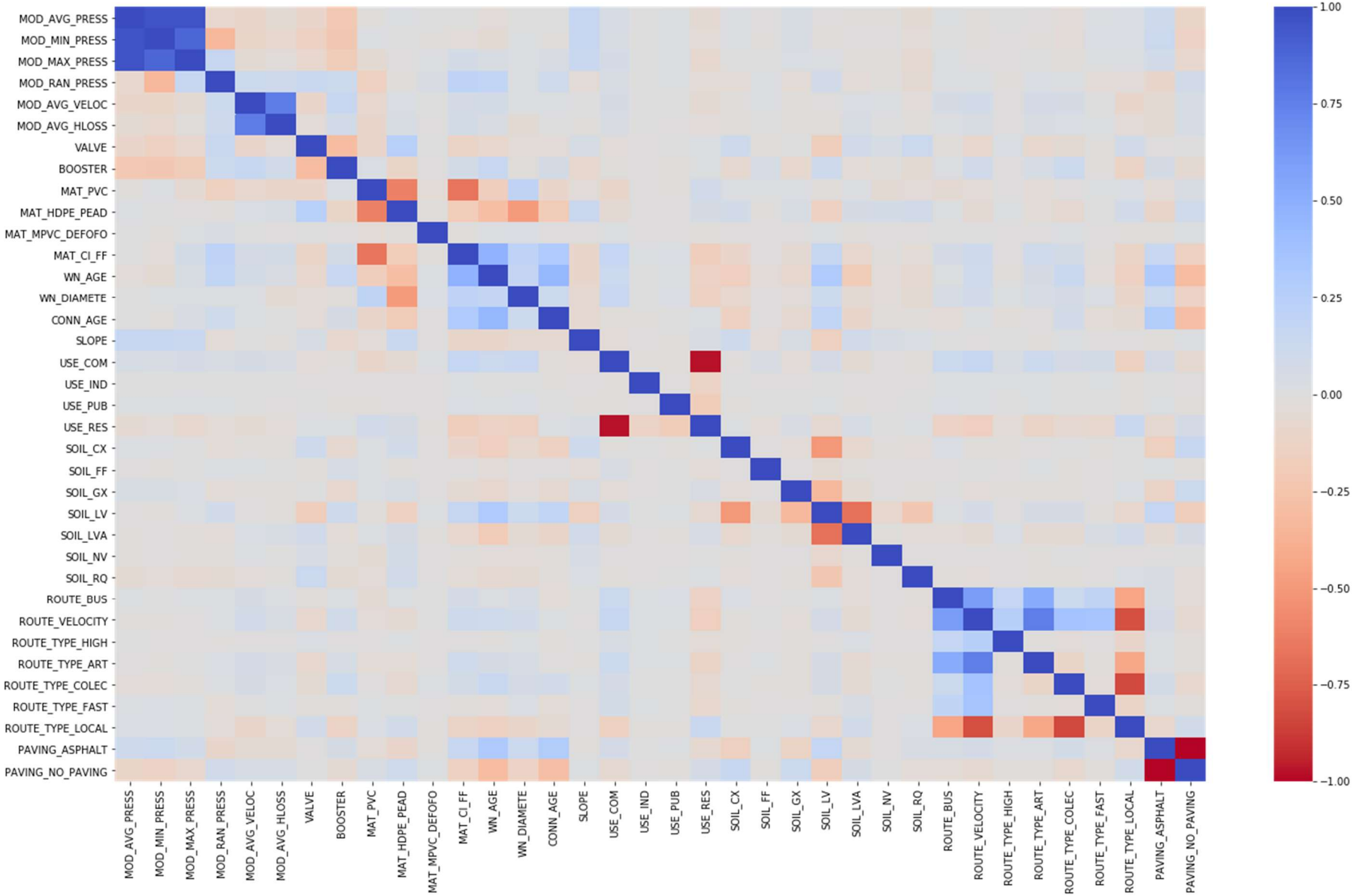


Figura 5.6 – Matriz de Correlação das variáveis que compõem o Banco de Dados.

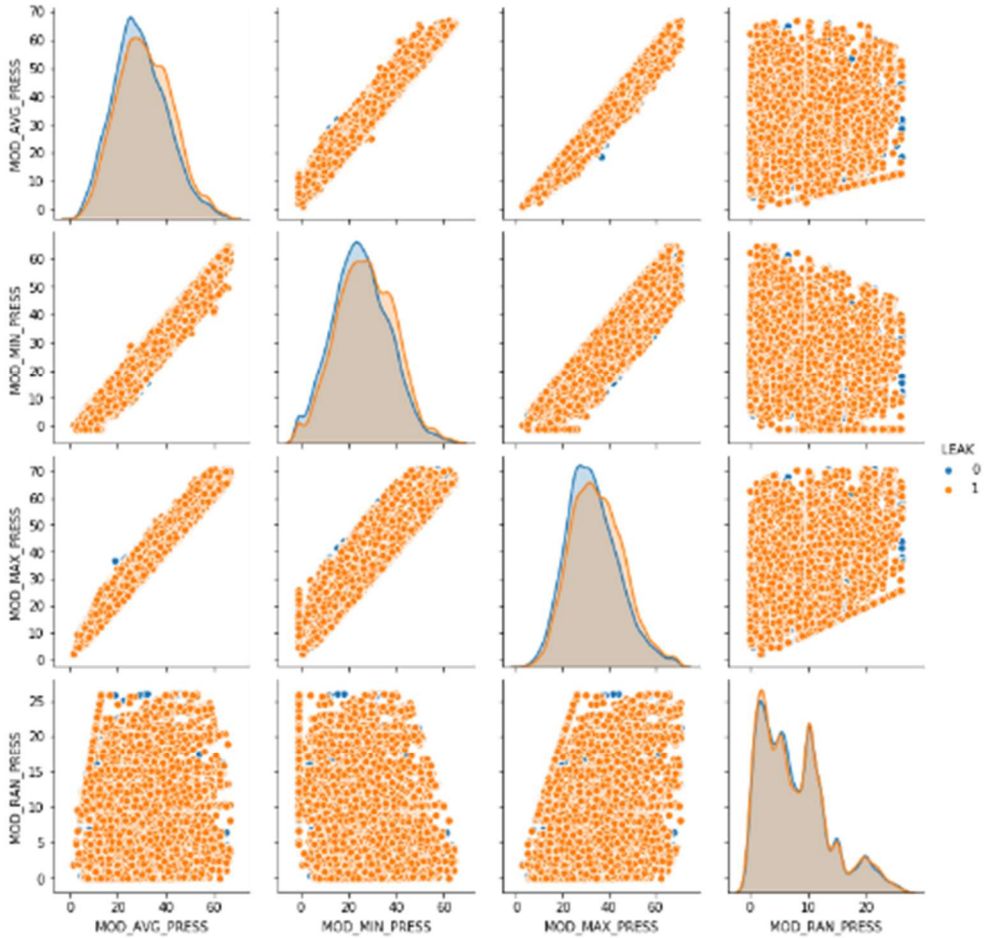


Figura 5.7 – Gráfico de pares para as variáveis de pressões médias, máximas, mínimas, e range de pressão.

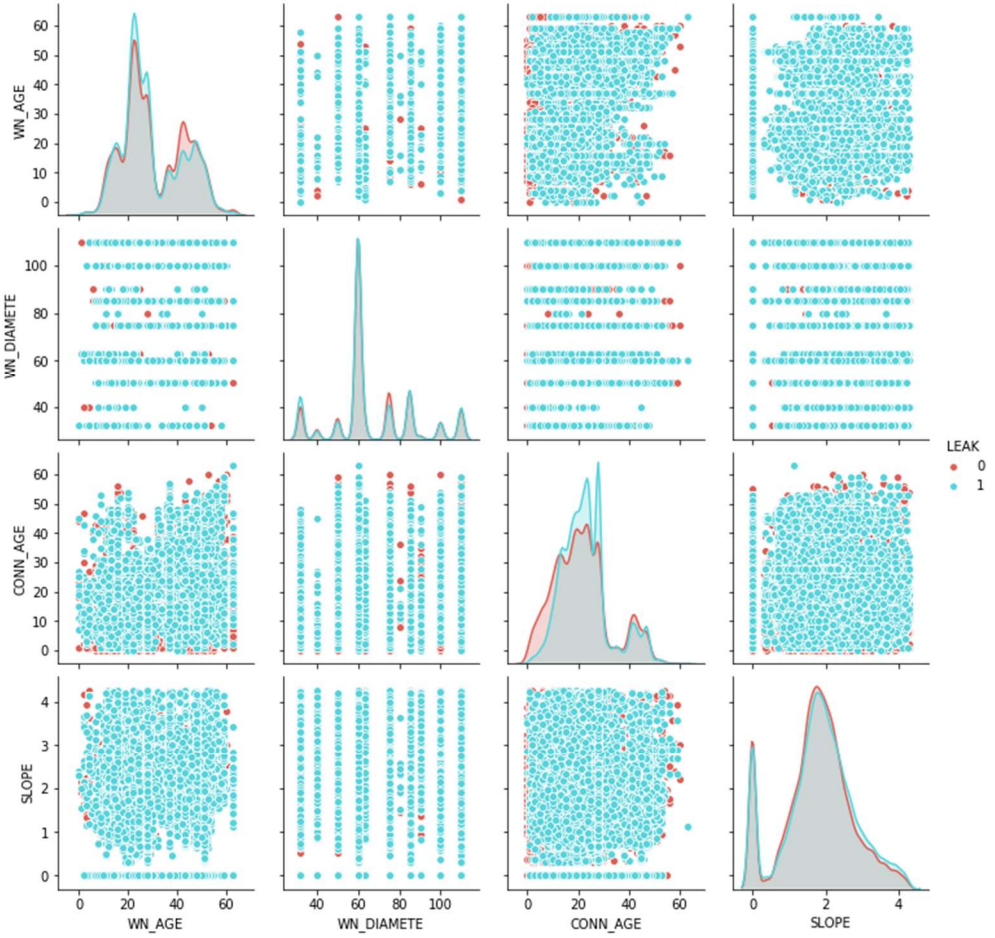


Figura 5.8 – Gráfico de pares para as variáveis de idade da rede, diâmetro da rede, idade da ligação de água, e declividade.

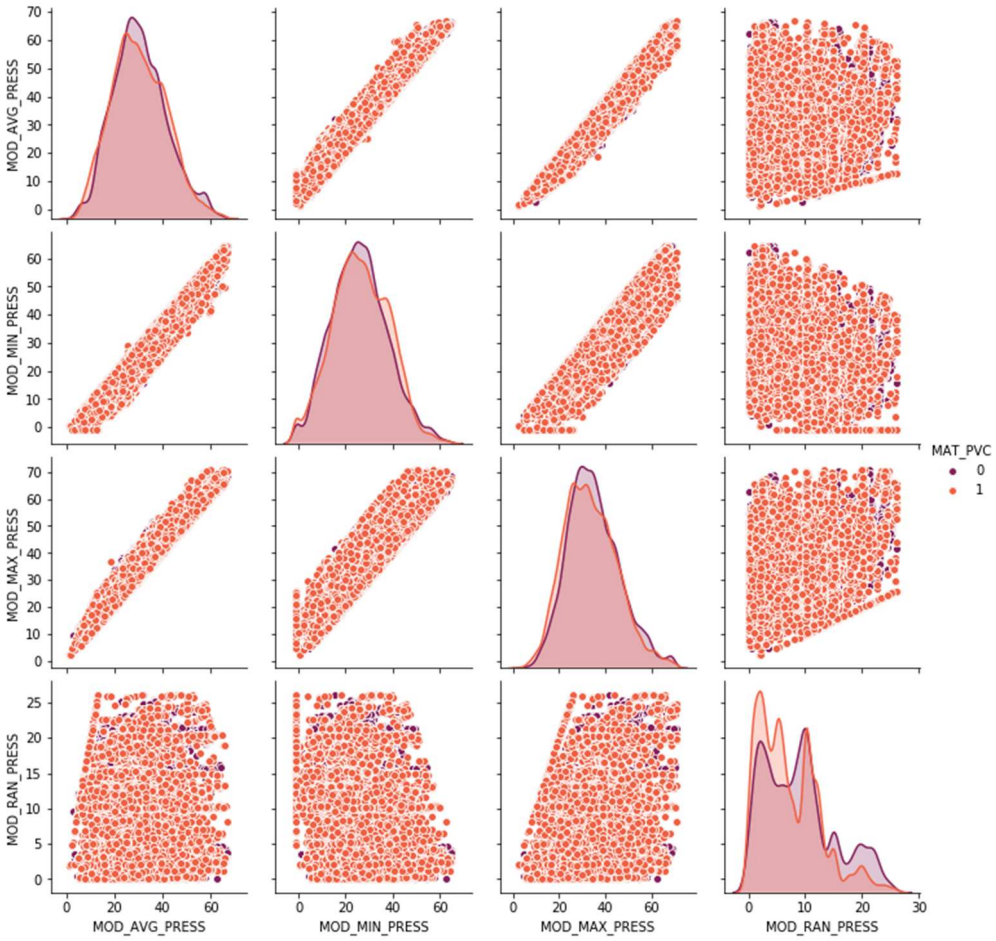


Figura 5.9 – Gráfico de pares para as variáveis de pressões médias, máximas, mínimas, e range de pressão; material PVC.

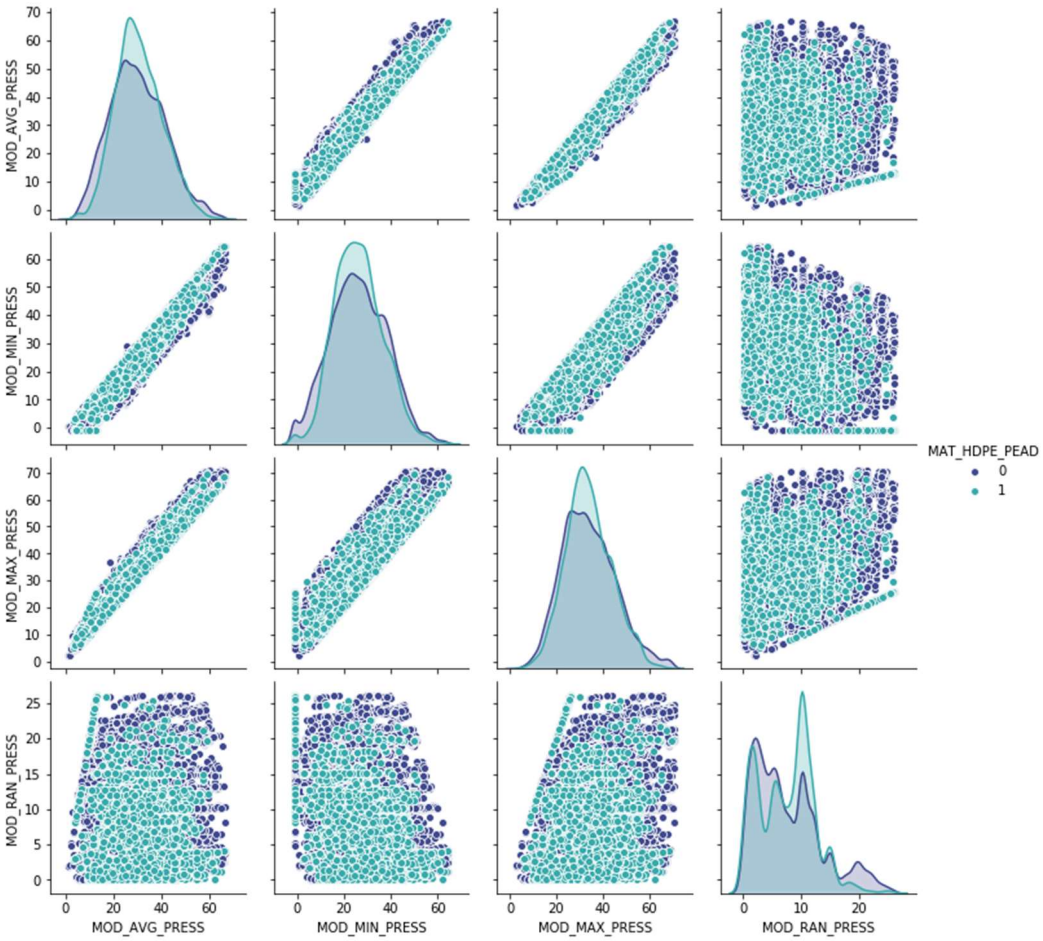


Figura 5.10 – Gráfico de pares para as variáveis de pressões médias, máximas, mínimas, e range de pressão; material PEAD.

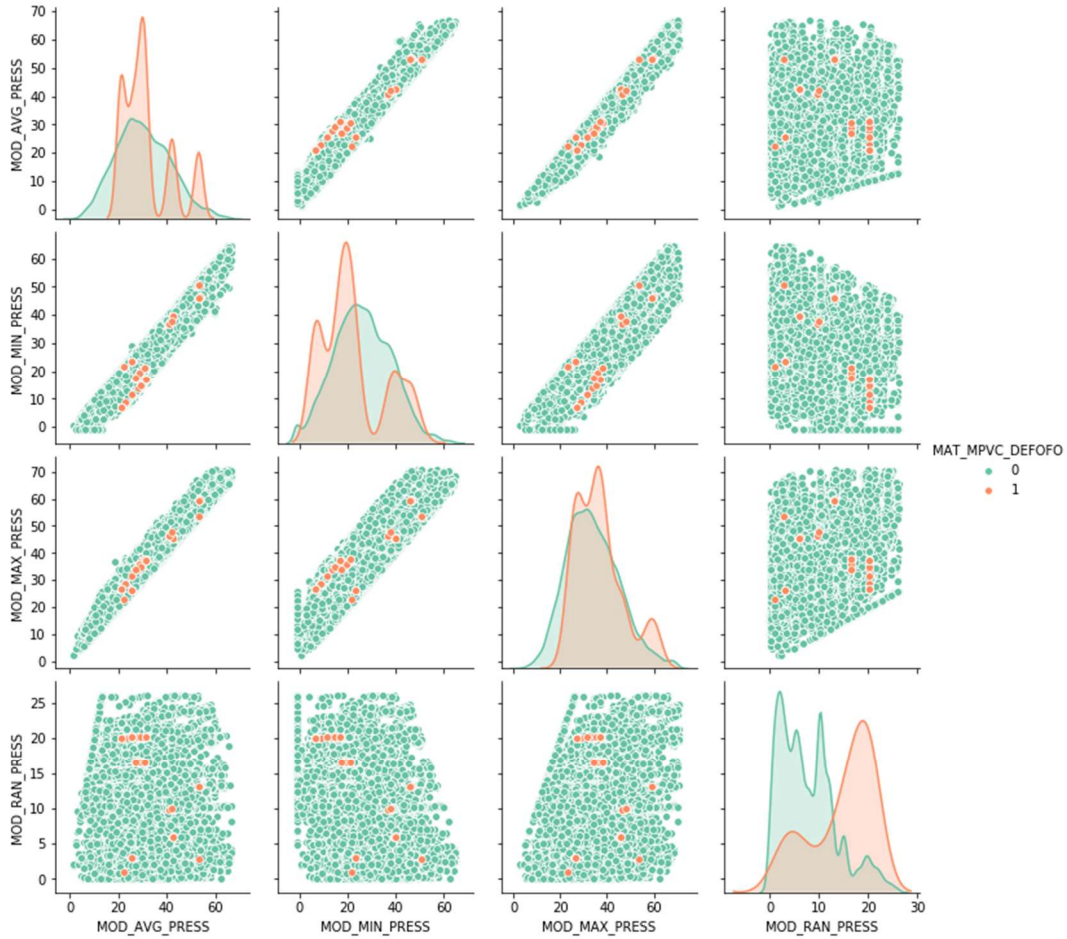


Figura 5.11 – Gráfico de pares para as variáveis de pressões médias, máximas, mínimas, e range de pressão; material DeFoFo.

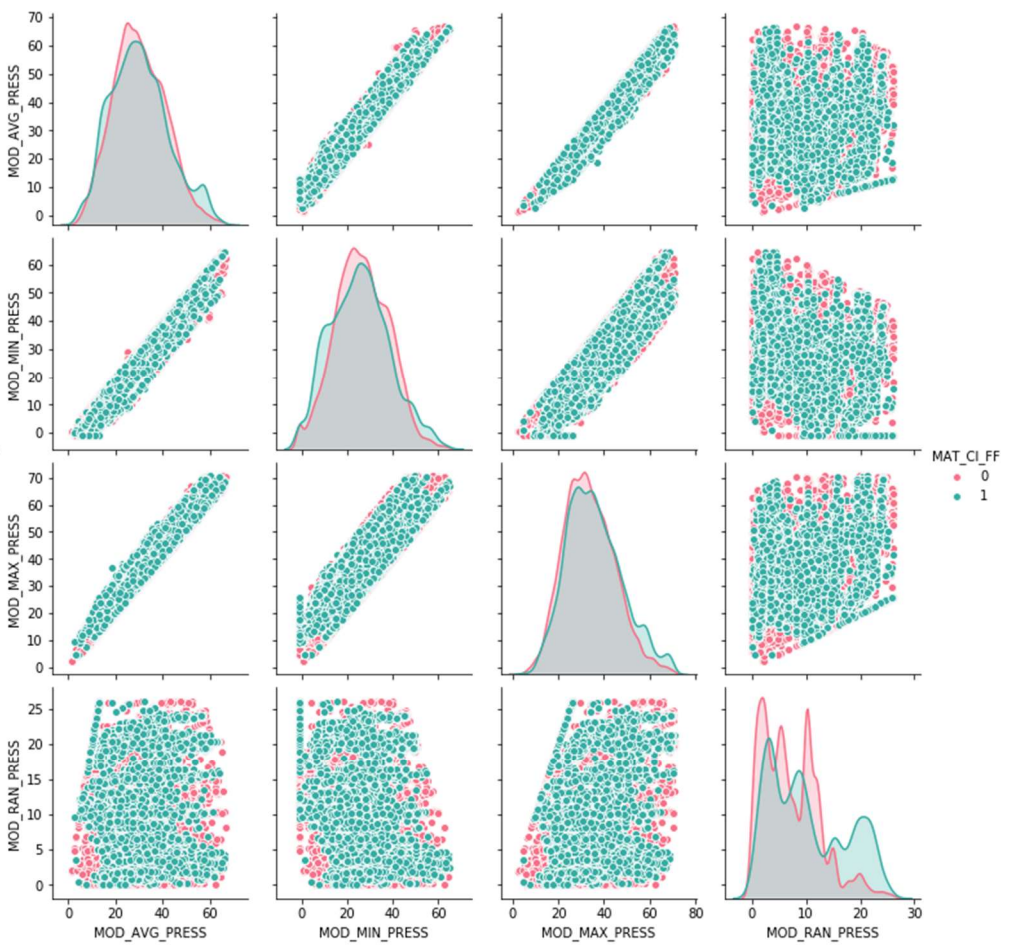


Figura 5.12 – Gráfico de pares para as variáveis de pressões médias, máximas, mínimas, e range de pressão; material FF.

Complementarmente, a Base de Dados também foi submetida à redução de dimensionalidade por três métodos diferentes: *Principal Component Analysis* (PCA); *t-distributed stochastic neighbor embedding* (t-SNE); e, *Truncated Singular Value Decomposition* (SVD). Os resultados mostram que a diferença entre as ligações que apresentaram vazamento visível reparado, e aquelas que não apresentaram, não é facilmente distinguível. Não há dois clusters prontamente identificáveis, visualmente pode-se constatar a sobreposição generalista da dispersão das classes por meio dos métodos ilustrados na Figura 5.13.

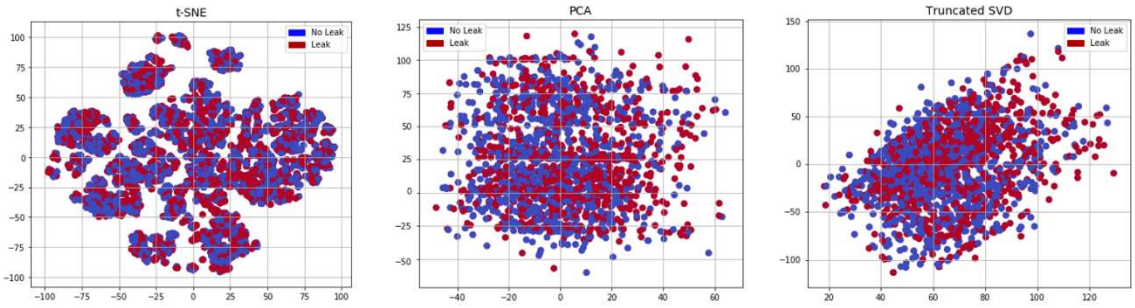


Figura 5.13 – Clusters utilizando Redução de Dimensionalidade aplicada na Base de Dados por meio de *Principal Component Analysis* (PCA); *t-distributed stochastic neighbor embedding* (t-SNE); e, *Truncated Singular Value Decomposition* (SVD).

As diferenças entre as classes inexistem ou são sutis. Portanto, o processamento da Base de Dados por modelos de Aprendizado de Máquina apresenta-se como ferramenta para investigação aprofundada sobre meios de predizer a classificação dos dados, revelando a relevância das variáveis preditoras. A Análise Exploratória dos Dados indica que a distinção entre as classes não é simplória, dado que a sobreposição de classes se mostra existente, conforme verificado. O próximo item discorre sobre a performance dos modelos aplicados, avaliando-se a acurácia para a classificação da Base de Dados.

5.2. PERFORMANCE DOS MODELOS

Os seguintes modelos de Aprendizado de Máquina foram aplicados à Base de Dados: *Linear Svm*, *Radial Svm*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, *Random Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*. A cada processamento, os dados foram submetidos à Validação Cruzada, utilizando-se 10 folders para cada iteração. 70% da Base de Dados foi aplicada para treinamento, 30% para teste. Todos os modelos foram aplicados, a cada vez, às redes atendidas pelos 12 Reservatórios Apoiados. Portanto, os dados foram processados 2.880 vezes (redes atendidas por 12 reservatórios diferentes; 12 modelos de Aprendizado de Máquina; 10 folders para Validação Cruzada; e, para cada reservatório, uma análise foi feita considerando os dados provenientes de modelo hidráulico, e outra sem esses dados).

A Base de Dados munida de dados hidráulicos obteve performances superiores, com exceção dos modelos *KNN*, *bagged KNN*, *Decision Tree*, e *Naive Bayes*, que tiveram performance pior com a inserção de mais variáveis. Esses modelos já não estavam entre os melhores modelos para classificar a base de dados, e não performaram melhor com o uso dos dados hidráulicos. Quatro modelos obtiveram performance superior a 57% em ambos os contextos, com e sem dados hidráulicos: *AdaBoost*, *XGBoost*, *Gradient Boosting*, e *Logistic Regression*.

O modelo com maior acurácia foi o *AdaBoost*, 59,23%, com uso do banco de dados completo, indicando que o uso de dados obtidos por meio de simulação hidráulica pode agregar valor e acurácia à predição. Sem dados hidráulicos, este modelo obteve acurácia de 57,88%. A Tabela 5.3 apresenta a performance obtida para os doze modelos aplicados, no cenário com e sem as variáveis hidráulicas.

Outra forma utilizada para explorar a performance dos modelos foi o agrupamento dos resultados por reservatório apoiado, para os casos com e sem dados provenientes da simulação hidráulica. As Figuras Figura 5.14, Figura 5.15 e Figura 5.16 apresentam a performance obtida por reservatório apoiado utilizando-se *boxplot* para a visualização.

A avaliação por reservatório identificou que diferentes redes de distribuição performaram acurácias diferentes, o que indica que a aplicação dos modelos de Aprendizado de

Máquina de forma regionalizada é pertinente. No geral, a utilização dos dados hidráulicos melhorou a acurácia das predições por reservatório, a exceção foram os reservatórios RAP.SSB.001 e RAP.SSB.002.

Tabela 5.3 – Acurácias médias obtidas por meio dos modelos *Linear Svm*, *Radial Svm*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, *Random Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*.

Acurácia média dos modelos de Aprendizado de Máquina [(VP+VN)/(VP+FP+VN+FN)]			
Modelo	**Sem dados hidráulicos**	**Modelo**	**Com dados hidráulicos**
AdaBoost	57,88%	*AdaBoost*	59,23%
Gradient Boosting	57,60%	*XGBoost*	57,88%
XGBoost	57,49%	*Gradient Boosting*	57,74%
Logistic Regression	57,01%	*Logistic Regression*	57,61%
Linear Svm	56,69%	*Radial Svm*	57,49%
Radial Svm	56,42%	*Linear Svm*	57,37%
KNN	56,23%	*Random Forest*	56,71%
bagged Decision Tree	56,08%	*bagged Decision Tree*	56,64%
Random Forest	56,04%	*KNN*	55,99%
bagged KNN	55,86%	*bagged KNN*	55,81%
Decision Tree	54,97%	*Decision Tree*	54,81%
Naive Bayes	53,92%	*Naive Bayes*	53,87%

O modelo de Aprendizado de Máquina *AdaBoost* apresentou melhor acurácia, assim, foi submetido à hiper-parametrização para verificação da possibilidade de incremento marginal de performance. A Tabela 5.4 resume os quatro cenários de aplicação do AdaBoost com e sem hiper-parametrização, com e sem os dados hidráulicos na Base de Dados. A hiper-parametrização demandou mais 280 rodadas de processamento da Base de Dados, totalizando 3.120 aplicações dos modelos aos dados.

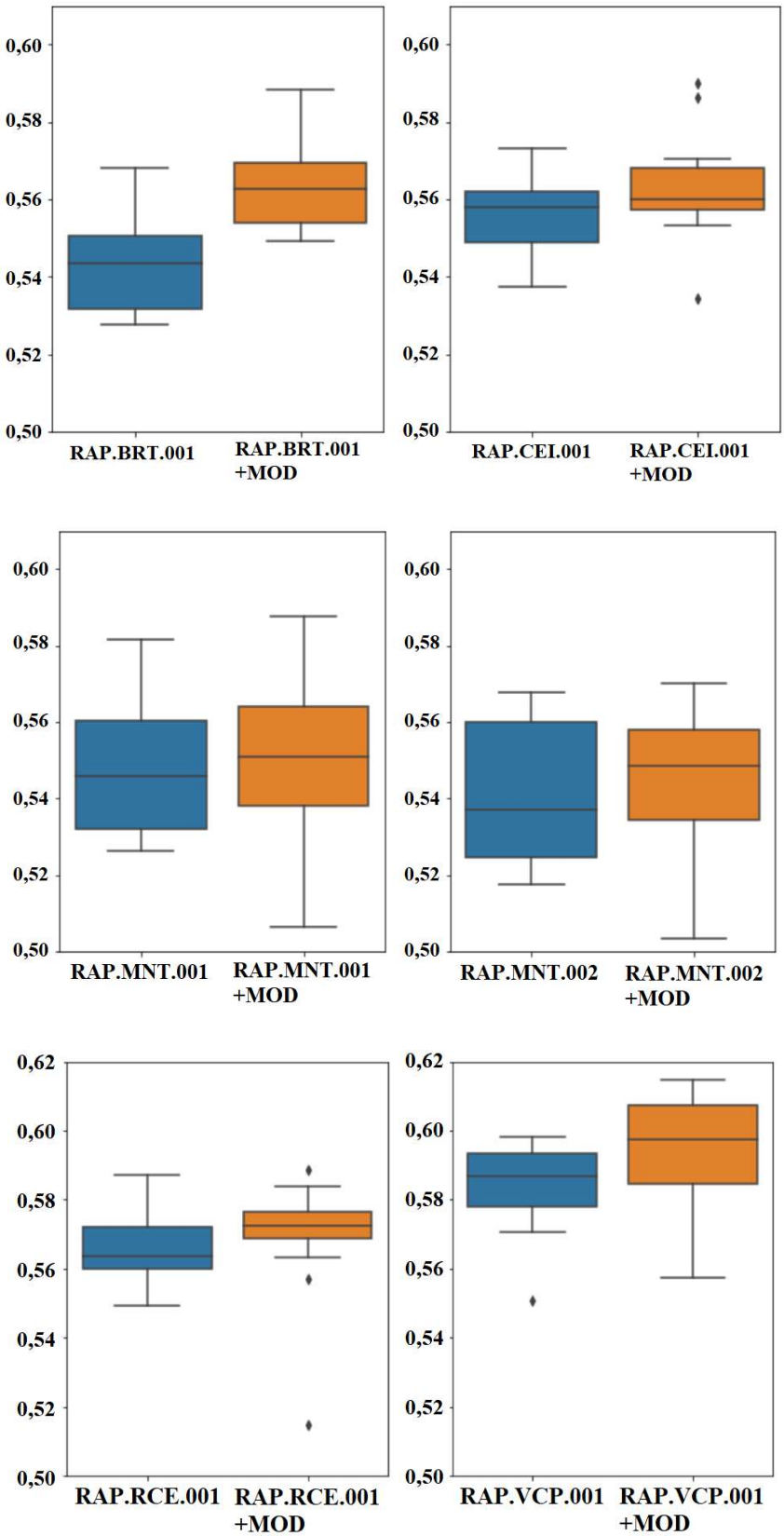


Figura 5.14 – *Boxplot* de performance obtida por meio dos modelos *Linear Svm*, *Radial Svm*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, Random *Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*, agrupados por reservatório apoiado (RAP.BRT.001, RAP.CEI.001, RAP.MNT.001, RAP.MNT.002, RAP.RCE.001, e RAP.VCP.001). "+MOD" refere-se aos resultados em que os dados de modelos hidráulicos foram utilizados.

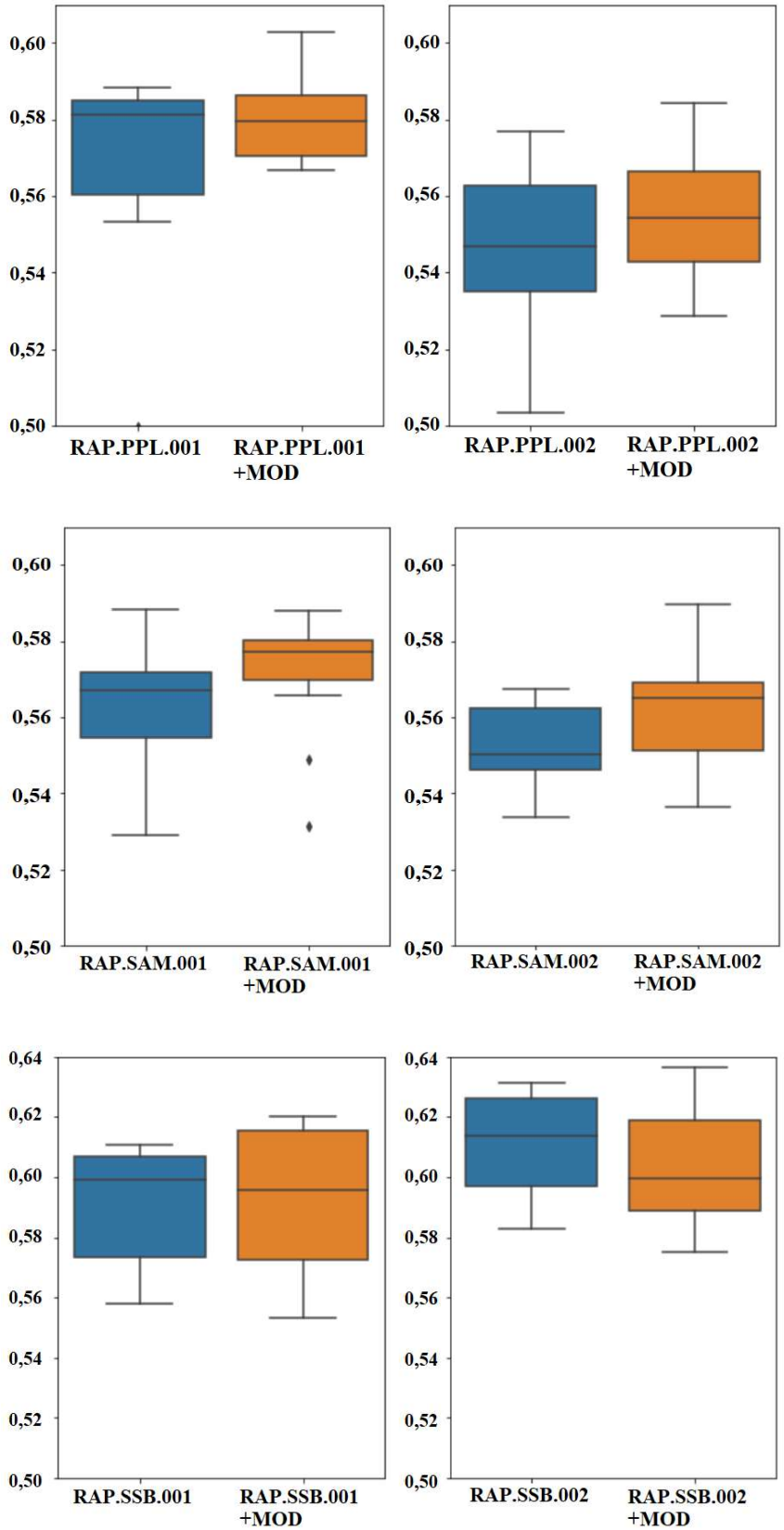


Figura 5.15 – *Boxplot* de performance obtida por meio dos modelos *Linear Svm*, *Radial Svm*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, Random *Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*, agrupados por reservatório apoiado (RAP.PPL.001, RAP.PPL.002, RAP.SAM.001, RAP.SAM.002, RAP.SSB.001, e RAP.SSB.002). "+MOD" refere-se aos resultados em que os dados de modelos hidráulicos foram utilizados.

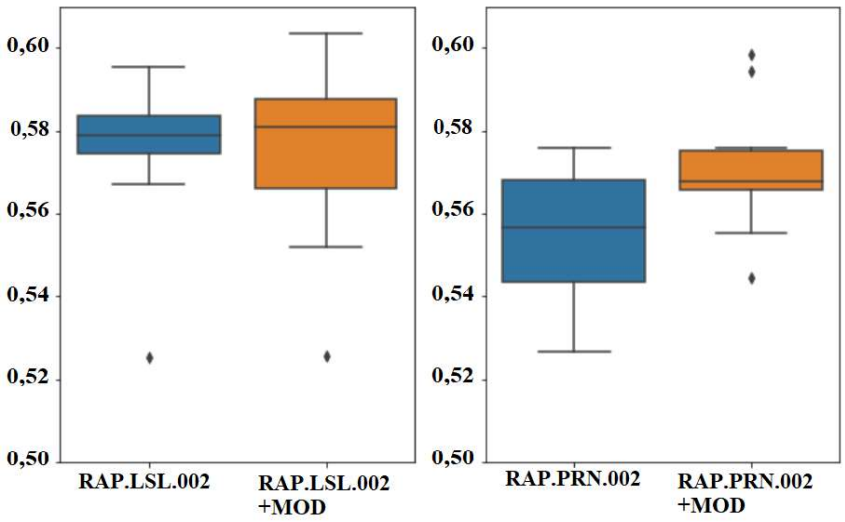


Figura 5.16 – *Boxplot* de performance obtida por meio dos modelos *Linear Svm*, *Radial Svm*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, *Random Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*, agrupados por reservatório apoiado (RAP.LSL.002 e RAP.PRN.002). "+MOD" refere-se aos resultados em que os dados de modelos hidráulicos foram utilizados.

Tabela 5.4 – Acurácias dos modelos *AdaBoost* e *AdaBoost* após *Hyper-Parameter Tuning*.

Acurácia média dos modelos *AdaBoost*			
Modelo	Sem dados hidráulicos	Modelo	Com dados hidráulicos
AdaBoost após *Hyper-Parameter Tuning*	58,51%	*AdaBoost* após *Hyper-Parameter Tuning*	59,70%
AdaBoost	57,88%	*AdaBoost*	59,23%

A performance dos modelos também pode ser apresentada por meio de *boxplot* agrupando-se as acurácias obtidas para os processamentos por reservatório. A Figura 5.17 apresenta um gráfico com as acurácias obtidas pelos doze modelos utilizados com dados hidráulicos, mais o *AdaBoost* com hiper-parametrização, sendo o modelo de referência, aleatório, com acurácia de 50%. A acurácia máxima obtida foi para o reservatório apoiado RAP.SSB.002, obtendo-se 63,63% de acerto por meio de processamento por após hiper-parametrização. Sete modelos e cinco reservatórios apoiados apresentaram acurácia superior a 60%, conforme apresentado na Tabela 5.5. Os reservatórios RAP.VCP.001 e RAP.SSB.002 foram os reservatórios em que mais modelos performaram acima de 60%. Por fim, o ganho de acurácia por meio da hiper-parametrização no modelo *AdaBoost* chegou a 4%, conforme apresentado na Tabela 5.6, indicando que a aplicação do método

pode representar melhoria ligeiramente significativa na acurácia do modelo em alguns casos.

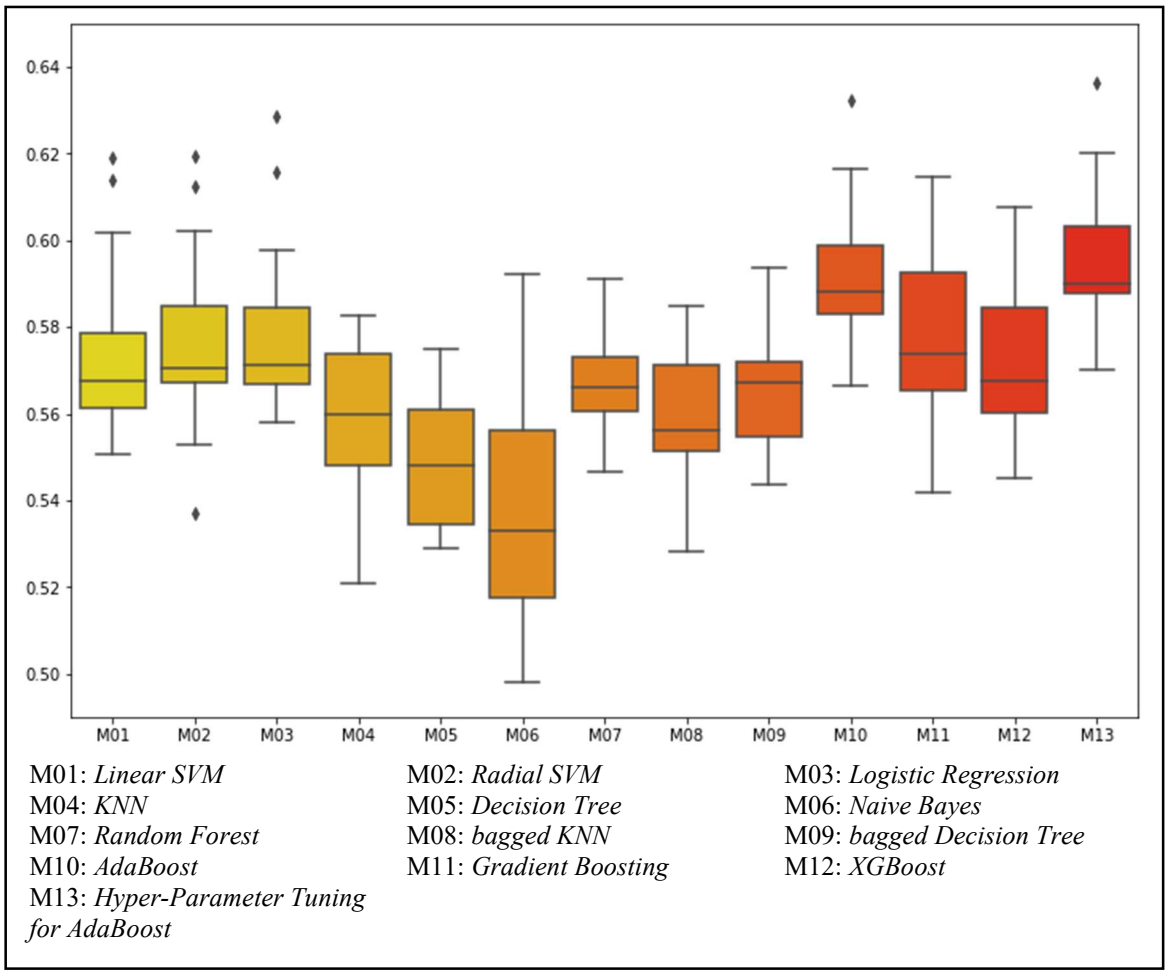


Figura 5.17 – Acurácias dos modelos aplicados.

Tabela 5.5 – Áreas de atendimento que obtiveram acurácia superior a 60%, com seus respectivos modelos.

Modelo	RAP.VCP.001	RAP.SSB.002	RAP.SSB.001	RAP.LSL.002	RAP.PPL.001
Linear Svm	60,19%	61,90%	61,38%	-	-
Radial Svm	60,21%	61,23%	61,93%	-	-
Logistic Regression	-	62,87%	61,56%	-	-
AdaBoost	60,84%	63,23%	61,65%	60,06%	-
Gradient Boosting	61,48%	-	-	-	-
XGBoost	60,76%	60,20%	-	-	-
AdaBoost após hiper-Parameter Tuning	61,23%	63,63%	62,02%	60,34%	60,27%

Tabela 5.6 – Acurácias dos modelos *AdaBoost* e *AdaBoost* após *Hyper-Parameter Tuning AdaBoost.*

Acurácia obtida por meio do modelo *AdaBoost* após *Hyper-Parameter Tuning*															
		RAP. SAM. 002	RAP. PRN. 002	RAP. BRT. 001	RAP. CEI. 001	RAP. VCP. 001	RAP. PPL. 001	RAP. SSB. 001	RAP. MNT. 002	RAP. LSL. 002	RAP. PPL. 002	RAP. SSB. 002	RAP. MNT. 001	RAP. SAM. 001	RAP. RCE. 001
AdaBoost após *Hyper-Parameter Tuning*	Sem dados hidráulicos	56,73%	57,58%	56,80%	57,30%	59,66%	58,76%	60,93%	56,00%	59,55%	57,70%	62,90%	58,15%	58,43%	58,73%
	Com dados hidráulicos	58,98%	59,85%	58,83%	59,00%	61,23%	60,27%	62,02%	57,00%	60,34%	58,43%	63,63%	58,75%	58,60%	58,88%
	Acréscimo de precisão	3,97%	3,95%	3,57%	2,97%	2,65%	2,56%	1,79%	1,79%	1,32%	1,26%	1,17%	1,03%	0,30%	0,26%

5.3. MATRIZES DE CONFUSÃO

No âmbito da aplicação de modelos de aprendizado de máquina, a avaliação da performance por meio de dados da Matriz de Confusão fornece informações aprofundadas sobre o desempenho dos algoritmos frente ao Banco de Dados processado. Assim, os resultados dos modelos aplicados (*Radial SVM, Linear SVM, KNN, Random Forest, Logistic Regression, Decision Tree, Naive Bayes, bagged KNN, bagged Decision Tree, Gradient Boosting, XGBoost, AdaBoost,* e *AdaBoost* após hiper-parametrização) em termos de Matriz de Confusão foram compilados em uma única tabela, prestada por meio da Tabela 5.7.

A acurácia, já discutida no item 5.2, trata sobre a precisão geral do modelo quanto aos acertos previstos. Em complemento, a Matriz de Confusão, item 3.9, provê informações sobre o quantitativo de registros classificados adequadamente como verdadeiro ou como falso, além daqueles classificados equivocadamente como verdadeiro ou como falso. Essas métricas são aplicadas no cálculo de novos indicadores, como a precisão e as taxas de sensibilidade e especificidade.

As melhores precisões para verdadeiro na Base de Dados, [1] *ramais que apresentaram vazamento visível reparado*, foram obtidas pelos modelos *XGBoost* e *AdaBoost*, 64,32 e 65,63%, respectivamente. O *AdaBoost* após hiper-parametrização obteve a melhor precisão para classificação de verdadeiros, 66,01%.

Embora o *AdaBoost* hiper-parametrizado apresente a melhor performance em termos de acurácia e precisão, a melhor precisão para classificação de falso na Base de Dados, [0] *ramais que* ***não*** *apresentaram vazamento visível reparado*, foi processada por meio de *Decision Tree* (58,73%), seguido do *Random Forest* (56,60%). A precisão para negativos, por meio do *AdaBoost* após hiper-parametrização, foi de 53,38%. Todavia, o *AdaBoost* após hiper-parametrização, por possuir a melhor acurácia, permanece com o melhor desempenho avaliando-se Taxa de Verdadeiro Positivo e Taxa de Verdadeiro Negativo, obtendo 58,67 e 61,04%, respectivamente.

Tabela 5.7 – Compilação das informações sobre Matriz de Confusão para os modelos de Aprendizado de Máquina aplicados.

Método	Verdadeiro Negativo (VN)	Falso Negativo (FN)	Falso Positivo (FP)	Verdadeiro Positivo (VP)	Acurácia (VP + VN)/(VP +VN+FP +FN)	Precisão VN/(VN + FN)	Precisão VP/(VP + FP)	Taxa de Verdadeiro Positivo (TVP) Sensibilidade [VP/(VP + FN)]	Taxa de Falso Negativo (TFN) 1 - Sensibilidade	Taxa de Verdadeiro Negativo (TVN) Especificidade [VN/(VN + FP)]	Taxa de Falso Positivo (TFP) 1 - Especificidade
Radial SVM	26,45%	23,59%	18,93%	31,04%	57,49%	52,86%	62,12%	56,82%	43,18%	58,29%	41,71%
Linear SVM	26,06%	24,22%	18,41%	31,32%	57,37%	51,82%	62,98%	56,39%	43,61%	58,60%	41,40%
KNN	27,14%	22,76%	21,25%	28,85%	55,99%	54,39%	57,58%	55,90%	44,10%	56,09%	43,91%
Random Forest	28,44%	21,81%	21,48%	28,27%	56,71%	56,60%	56,83%	56,44%	43,56%	56,98%	43,02%
Logistic Regression	27,68%	22,34%	20,05%	29,93%	57,61%	55,33%	59,88%	57,26%	42,74%	57,99%	42,01%
Decision Tree	29,40%	20,66%	24,53%	25,41%	54,81%	58,73%	50,88%	55,16%	44,84%	54,51%	45,49%
Naive Bayes	23,56%	26,48%	19,65%	30,31%	53,87%	47,08%	60,67%	53,37%	46,63%	54,52%	45,48%
bagged KNN	25,30%	24,37%	19,81%	30,52%	55,81%	50,93%	60,63%	55,60%	44,40%	56,08%	43,92%
bagged Decision Tree	26,46%	24,12%	19,24%	30,18%	56,64%	52,31%	61,07%	55,58%	44,42%	57,89%	42,11%
Gradient Boosting	26,57%	22,77%	19,49%	31,17%	57,74%	53,85%	61,53%	57,78%	42,22%	57,69%	42,31%
XGBoost	25,51%	24,17%	17,95%	32,36%	57,88%	51,35%	64,32%	57,24%	42,76%	58,70%	41,30%
AdaBoost	26,46%	23,60%	17,16%	32,78%	59,23%	52,85%	65,63%	58,14%	41,86%	60,65%	39,35%
AdaBoost Hiper-param.	26,66%	23,28%	17,01%	33,05%	59,70%	53,38%	66,01%	58,67%	41,33%	61,04%	38,96%

5.4. VARIÁVEIS PREDITORAS À FALHA

Em continuidade à avaliação dos resultados proporcionados pela aplicação dos modelos de Aprendizado de Máquina à Base de Dados, a relevância das variáveis preditoras à falha foi apurada por reservatório apoiado, considerando os cenários sem e com dados provenientes dos modelos hidráulicos.

A sequência de Figuras entre Figura 5.18 e Figura 5.31 apresenta as variáveis com maior relevância, para os cenários citados e para os reservatórios RAP.BRT.001, RAP.CEI.001, RAP.LSL.002, RAP.MNT.001, RAP.MNT.002, RAP.PPL.001, RAP.PPL.002, RAP.PRN.002, RAP.RCE.001, RAP.SAM.001, RAP.SAM.002, RAP.SSB.001, RAP.SSB.002, e RAP.VCP.001. À esquerda, em cada figura, são apresentadas as variáveis processadas na Base de Dados sem o uso das condições operacionais da rede obtidas por meio de simulação hidráulica. À direita, os resultados obtidos integrando-se as variáveis hidráulicas à Base de Dados. Tais figuras ilustram o desempenho para as variáveis com maior relevância para os seguintes modelos aplicados: *AdaBoost*, *XGBoost*, *Gradient Boost* e *Random Forest*. Os três primeiros modelos foram utilizados nessas figuras dado que são os modelos com melhor acurácia e precisão para valores verdadeiros. O *Random Forest*, embora não tenha performado entre os melhores modelos em termos de acurácia, foi o segundo melhor modelo em precisão para valores negativos; então, optou-se por incluí-lo na apresentação gráfica para permitir comparação sobre a forma como a relevância foi obtida entre as variáveis. O modelo com melhor precisão para valores negativos foi *Decision Tree*, mas a acurácia para esse modelo foi inferior a 55%, motivo pelo qual o *Random Forest* foi apresentado.

Sem a aplicação dos dados hidráulicos, obtêm-se destaque para as variáveis idade da conexão/ligação de água e declividade do terreno sob a rede de distribuição que provê derivação à ligação. Em sequência, idade e diâmetro da rede performam como variáveis relevantes. Em termos gerais, quanto à Base de Dados processada com as variáveis hidráulicas, os principais preditores à falha são: idade da conexão, pressão máxima operacional, e variação de pressão (range). As Tabelas Tabela 5.8 e Tabela 5.9 apresentam uma síntese das variáveis preditoras, uma soma entre os resultados obtidos por meio dos modelos *AdaBoost*, *XGBoost*, *Gradient Boost* e *Random Forest*.

A aplicação dos modelos por área de atendimento, delimitada por reservatório apoiado, se mostra pertinente, uma vez que há variância entre a relevância das variáveis por região. As ligações atendidas por meio dos reservatórios RAP.LSL.002 e RAP.BRT.001 são mais suscetíveis à vazamentos quanto assentadas sob solo do tipo cambissolo háplico. As ligações constantes do RAP.BRT.001 possuem como segunda variável preditora a variação de pressão, enquanto a média dos resultados aponta a pressão máxima. Na área atendida pelo RAP.CEI.001, as ligações provenientes de tubulações em ferro fundido são mais suscetíveis ao vazamento visível, em detrimento das tubulações em materiais plásticos.

A inserção das variáveis hidráulicas proporcionou em alguns casos que variáveis sem expressão no processamento sem essas variáveis obtivessem relevância. Após a inserção das variáveis hidráulicas, para o RAP.MNT.001, a existência de VRP e tubos em ferro fundido obtiveram maior relevância. No caso do RAP.PRN.002, a existência de VRP passou, inclusive, a obter o maior índice quanto à predição da falha.

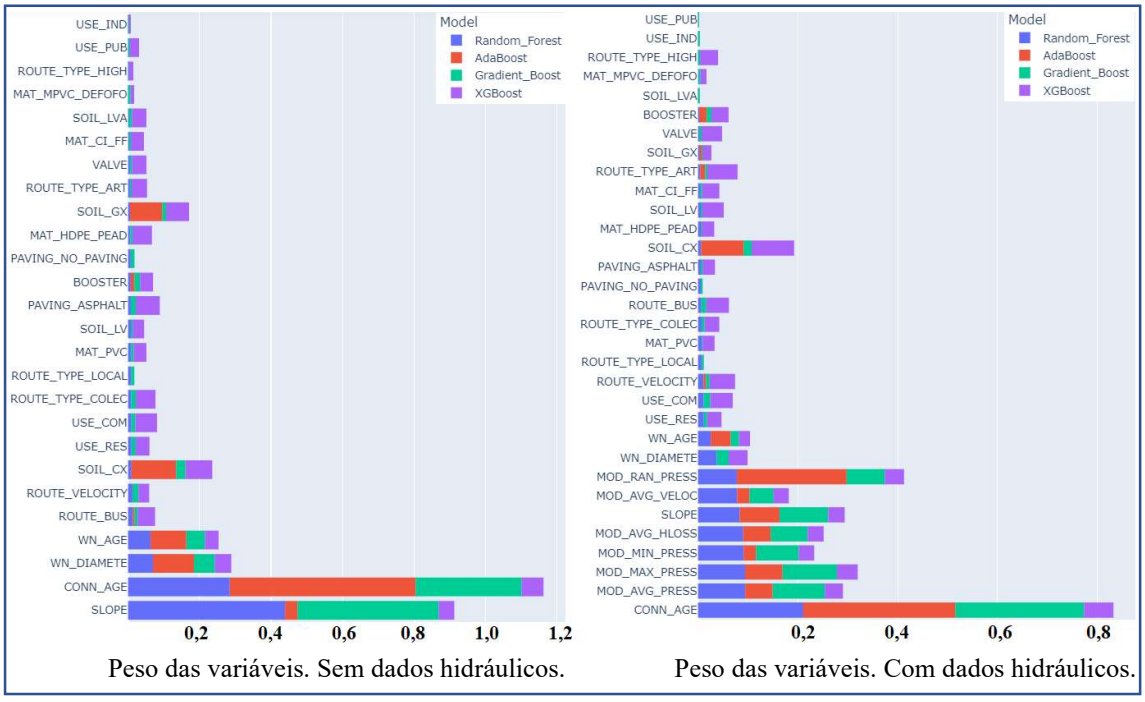
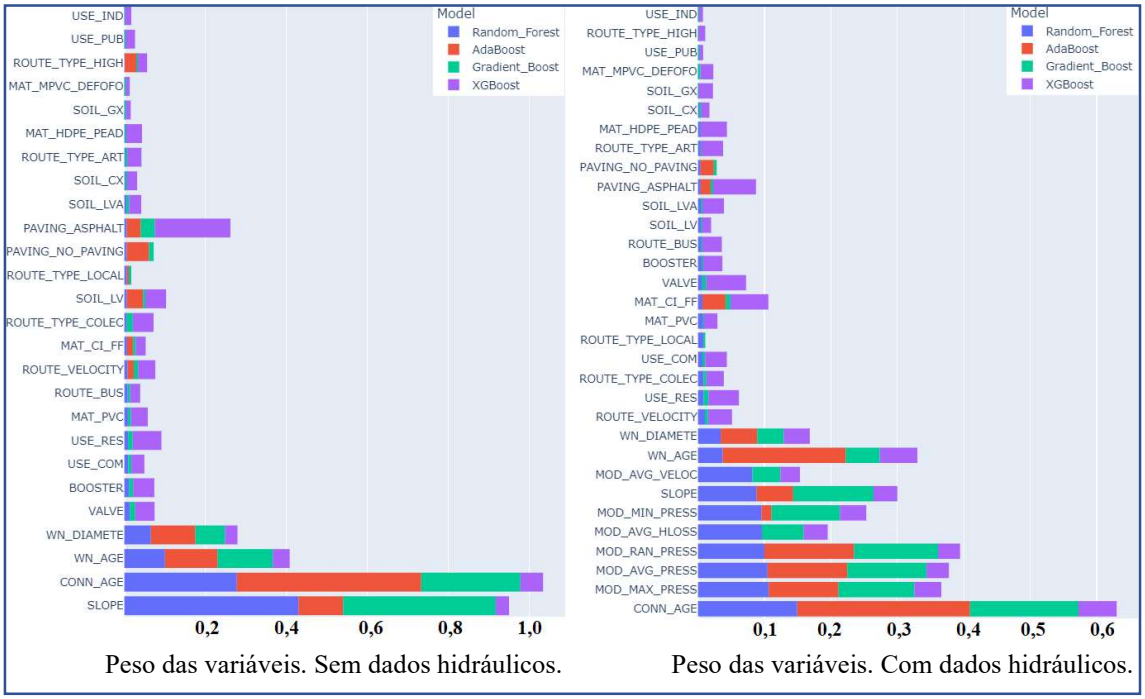
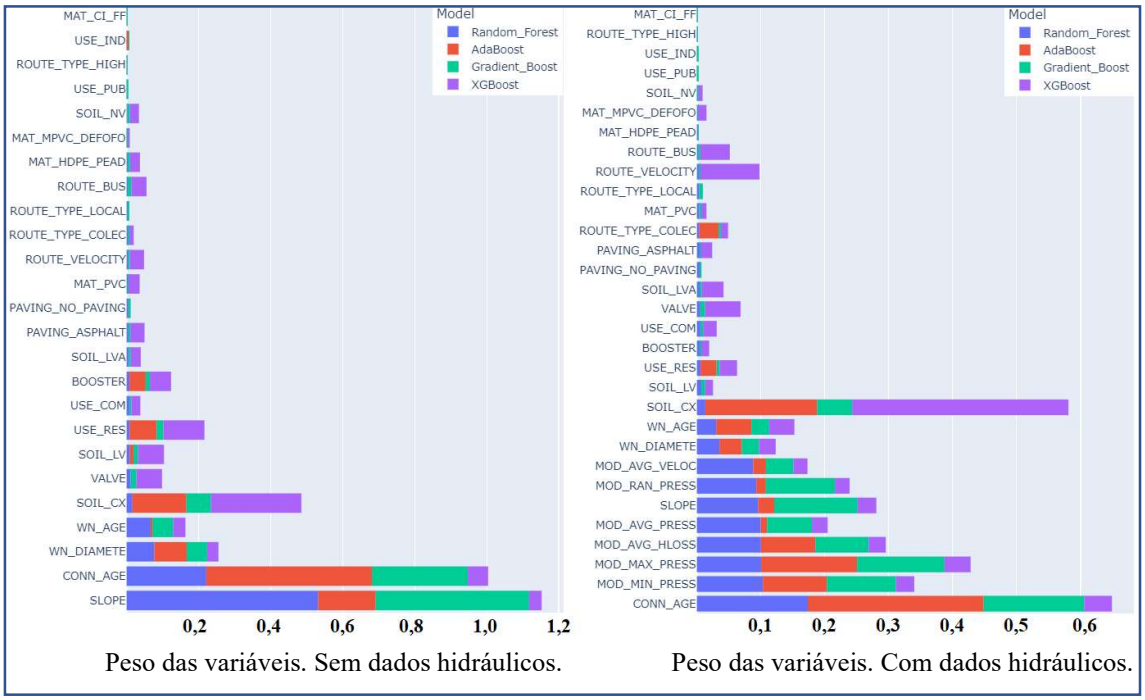


Figura 5.18 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.BRT.001.



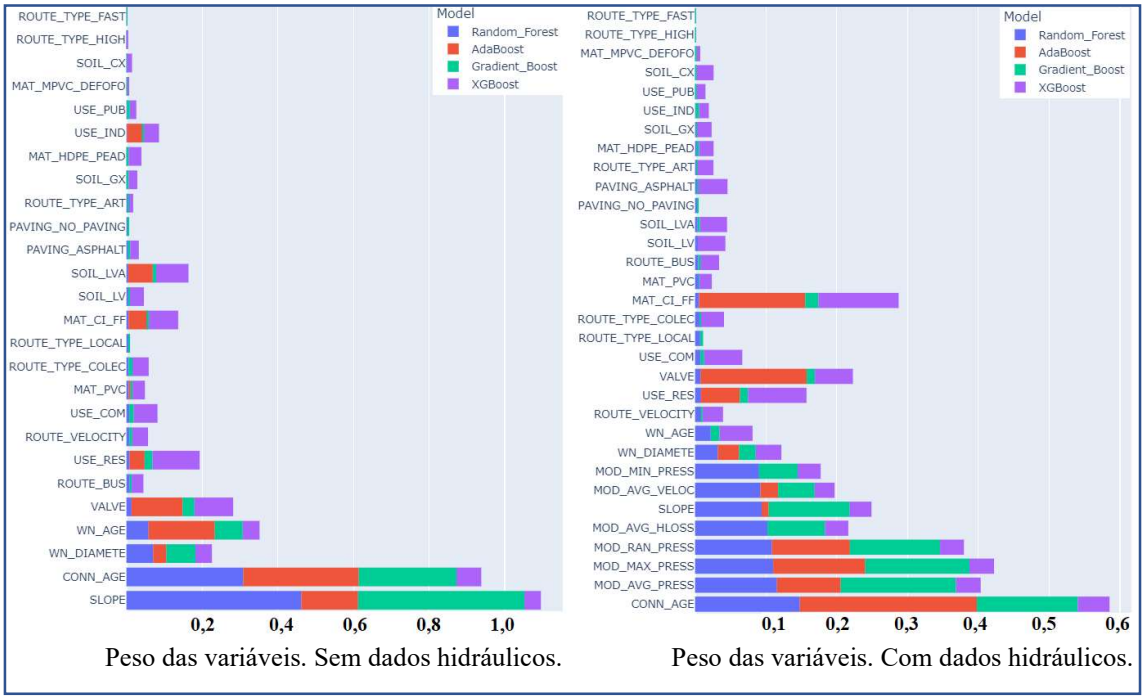
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.19 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.CEI.001.



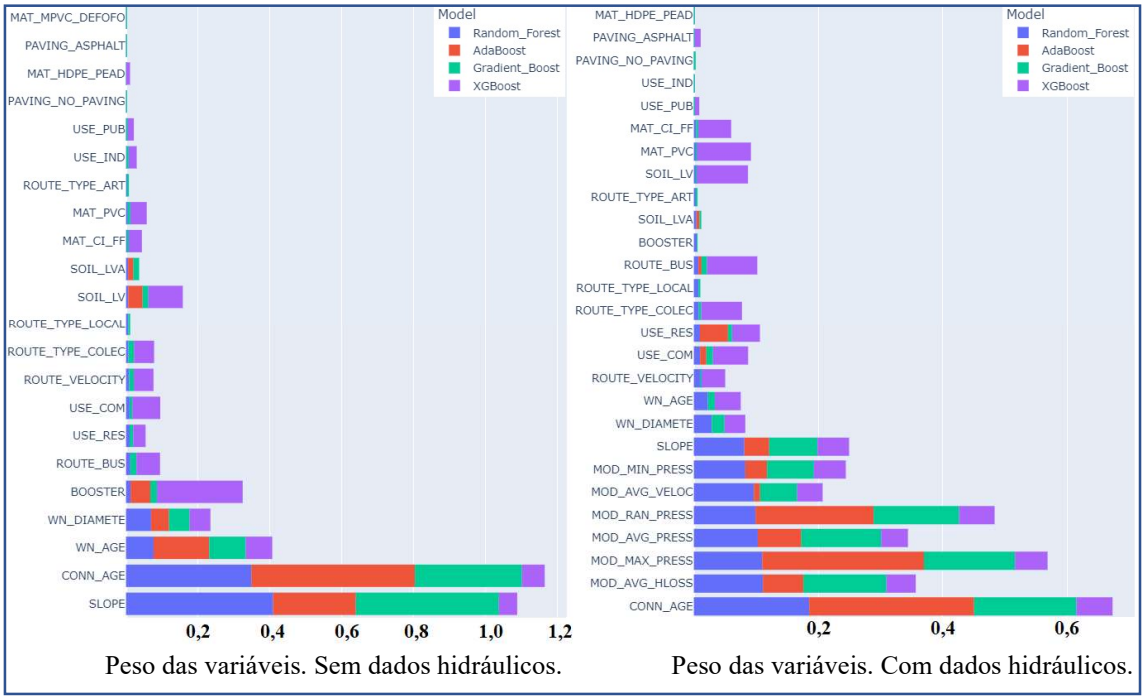
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.20 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.LSL.002.



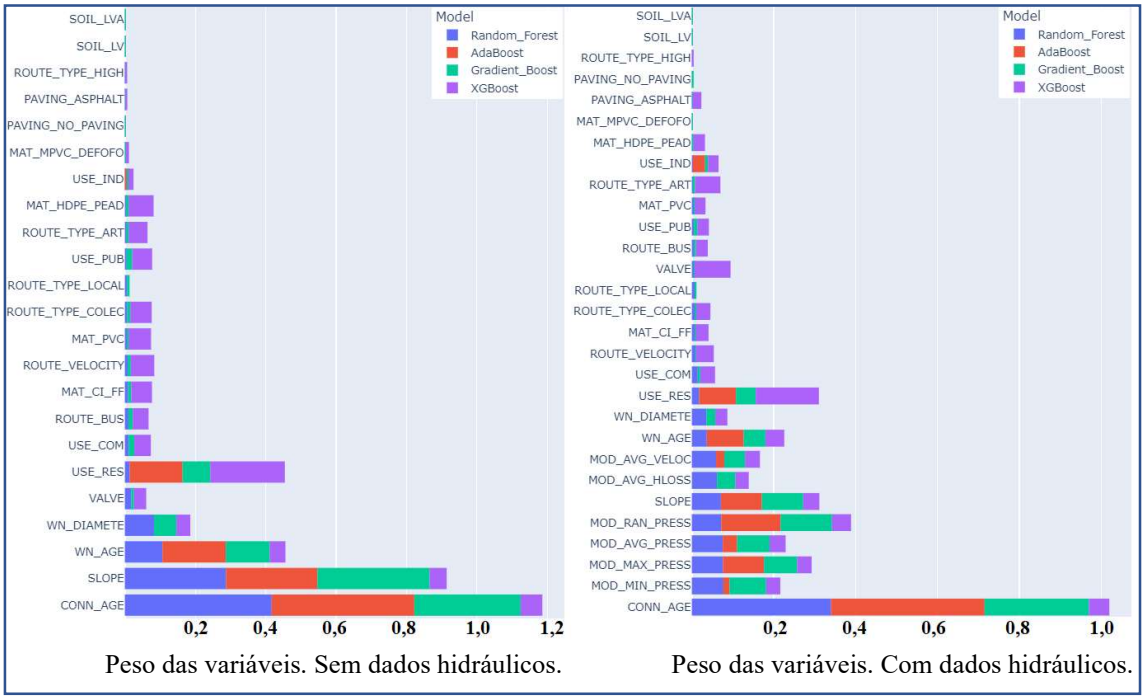
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.21 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.MNT.001.



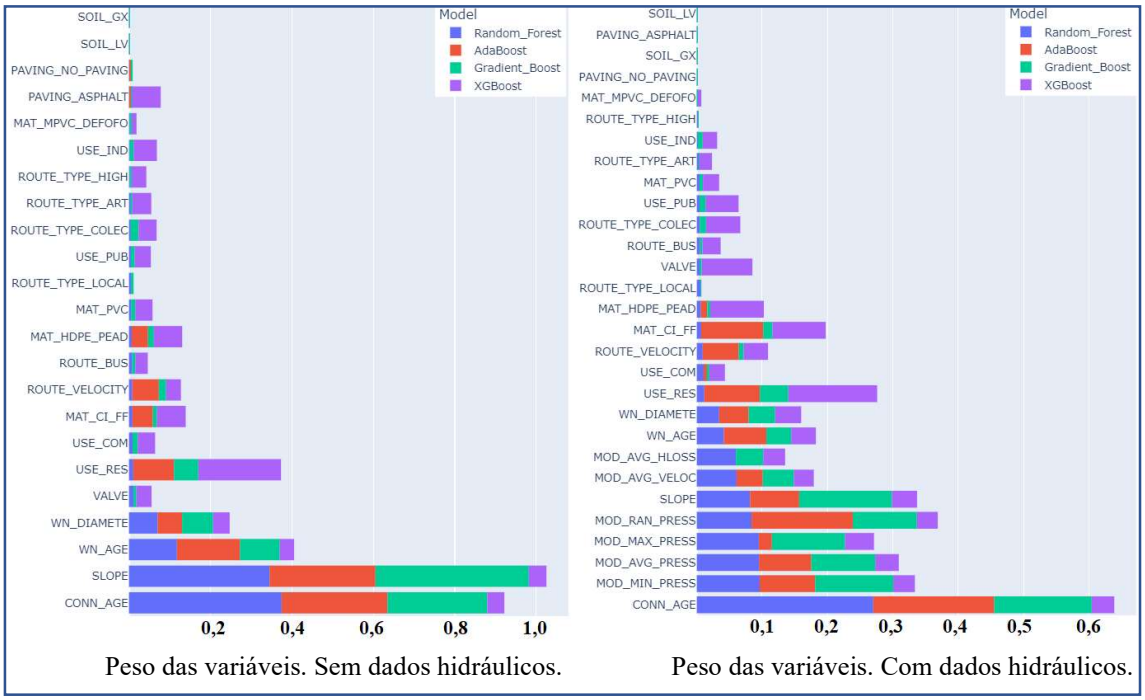
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.22 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.MNT.002.



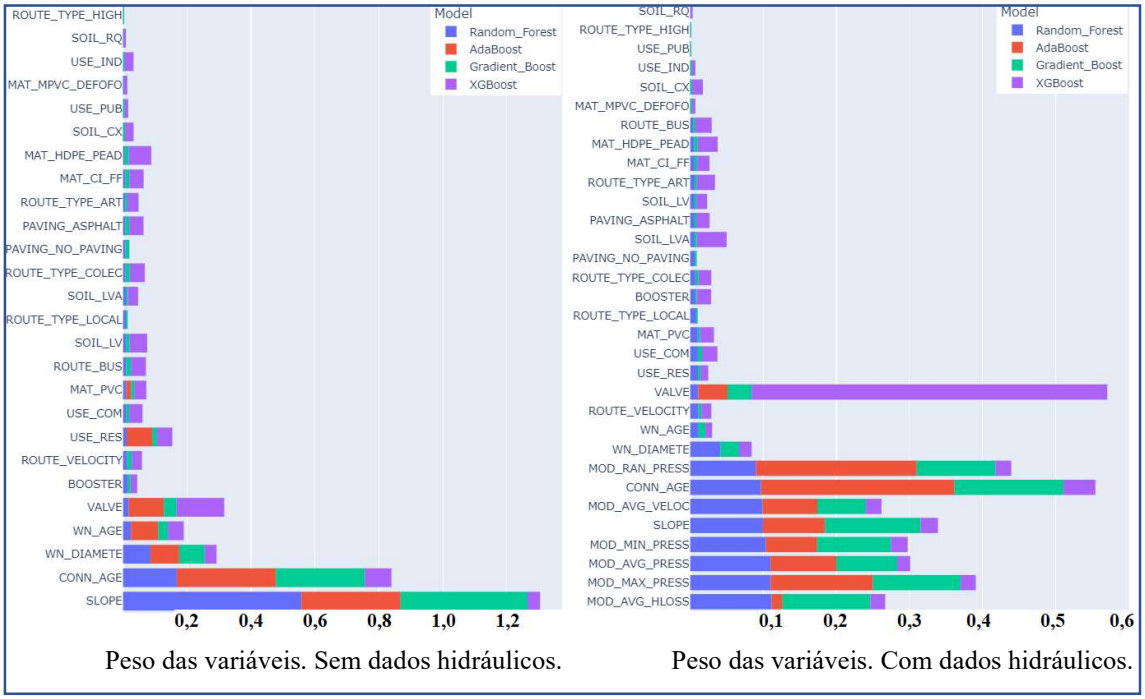
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.23 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.PPL.001.



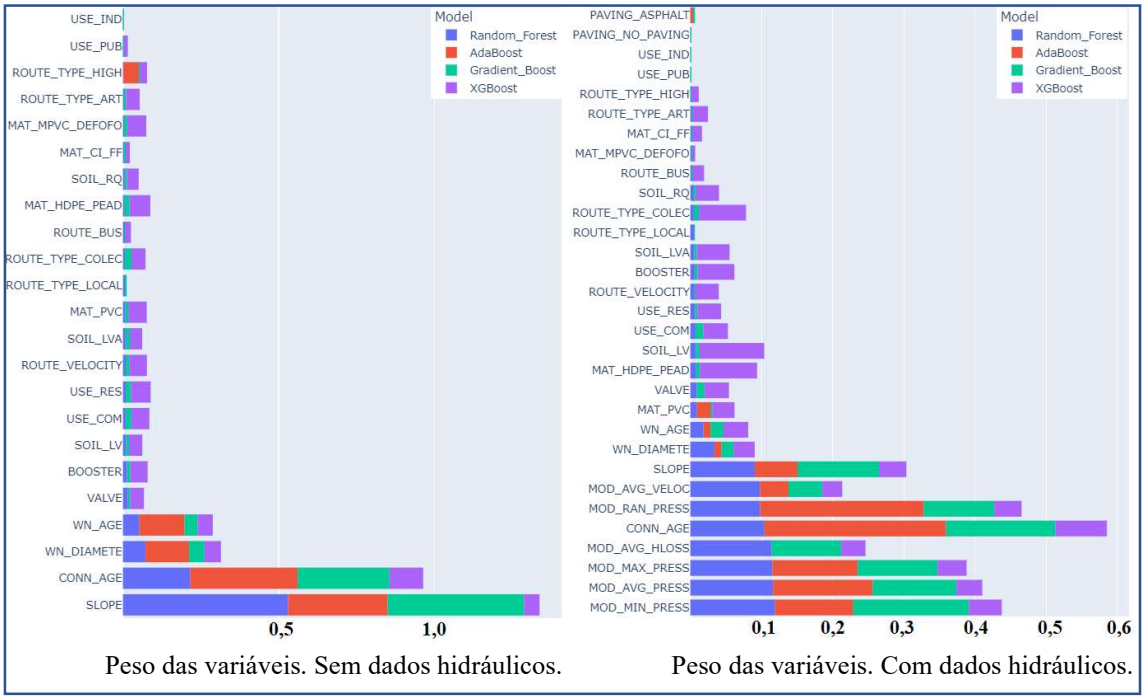
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.24 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.PPL.002.



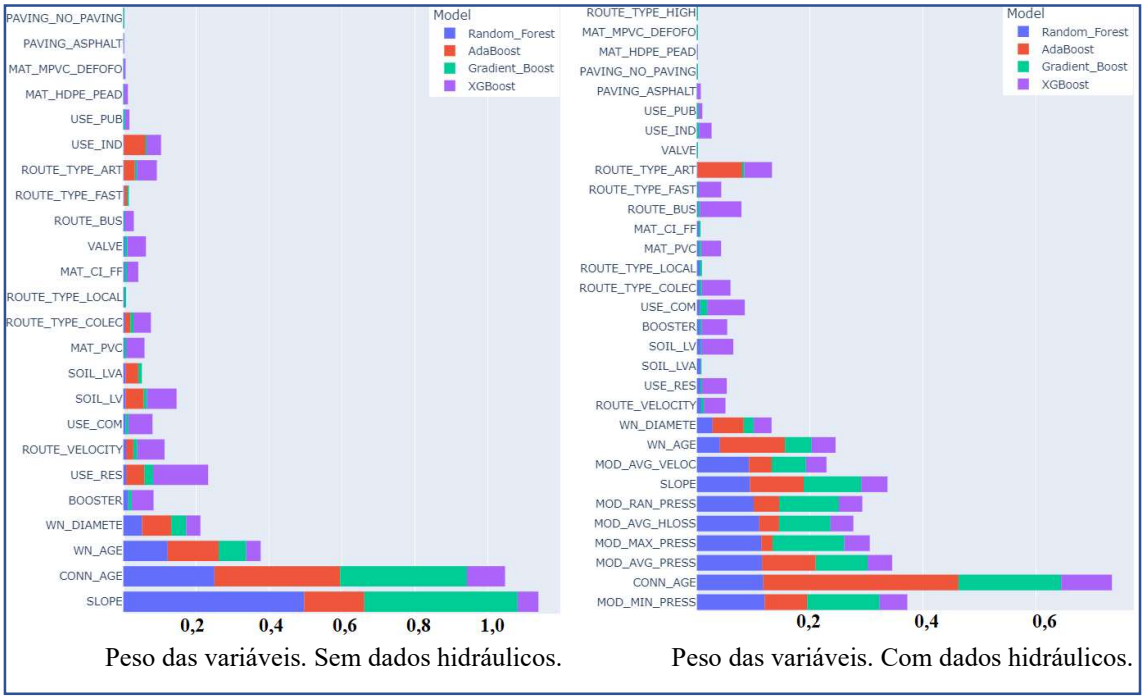
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.25 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.PRN.002.



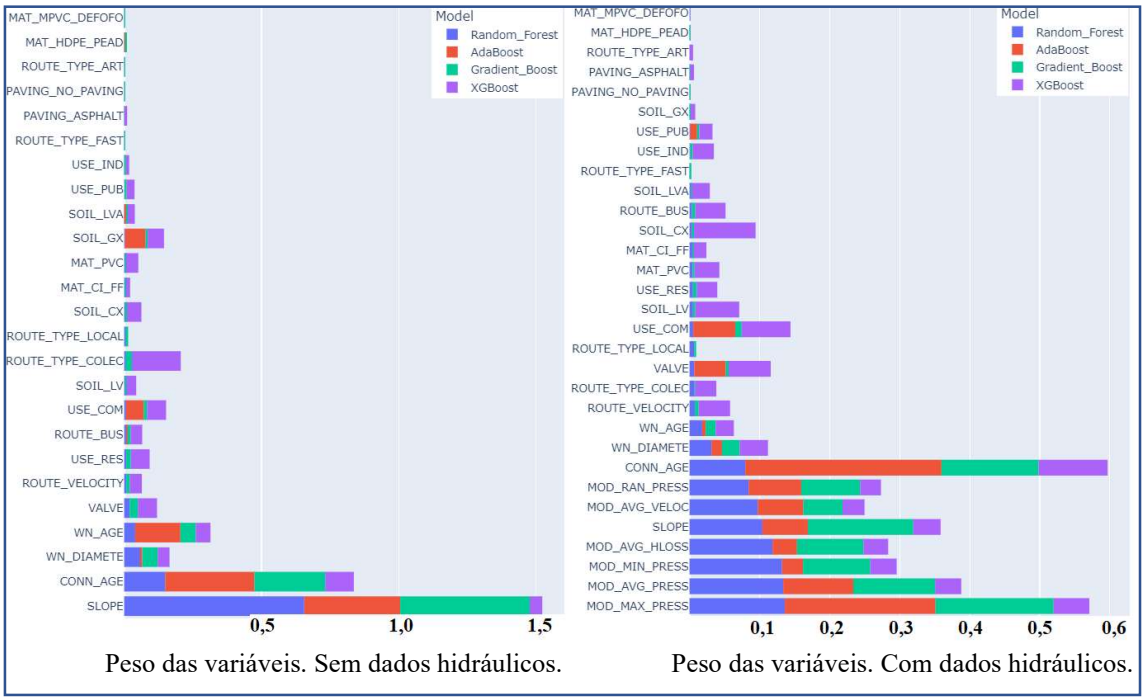
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.26 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.RCE.001.



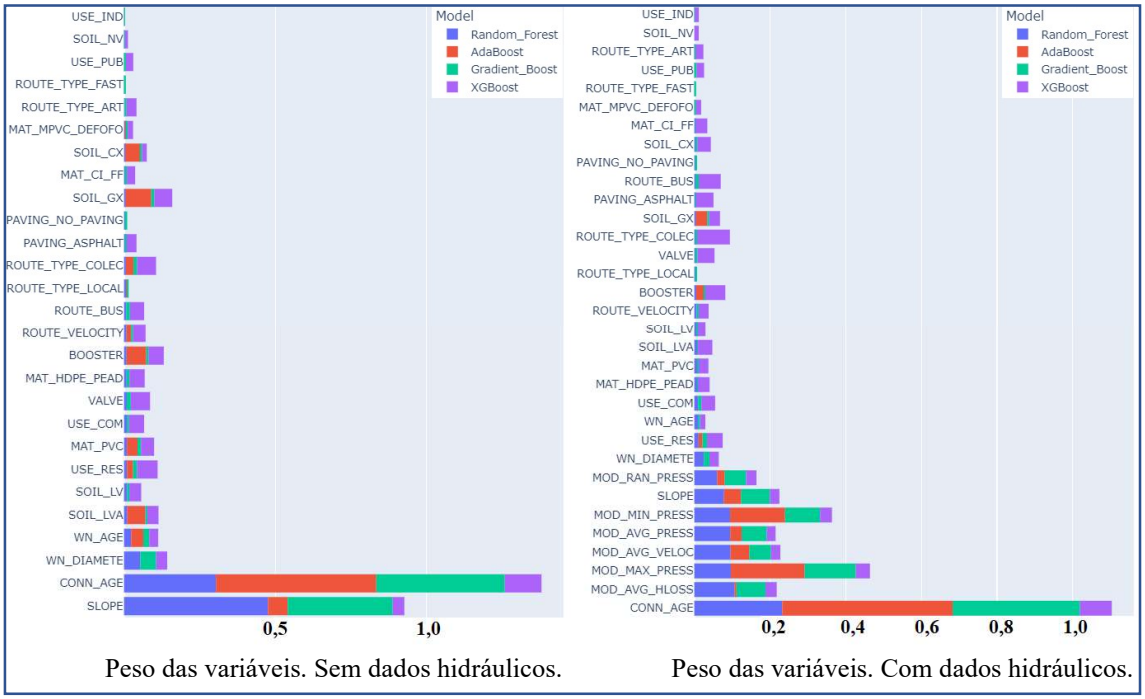
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.27 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.SAM.001.



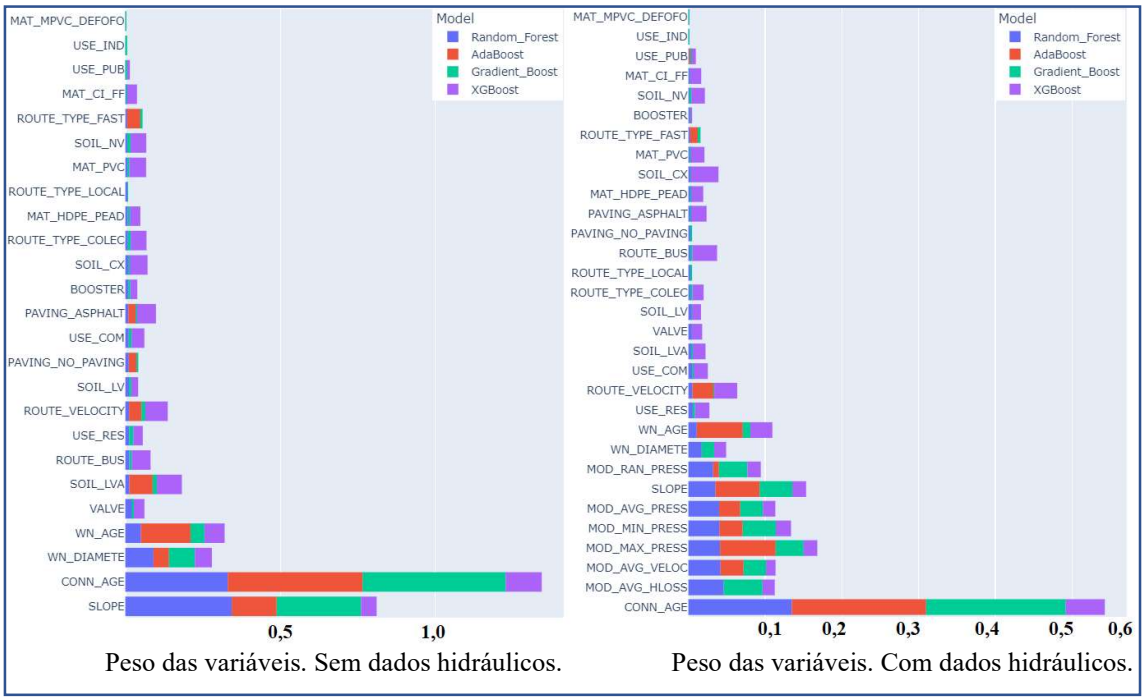
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.28 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.SAM.002.



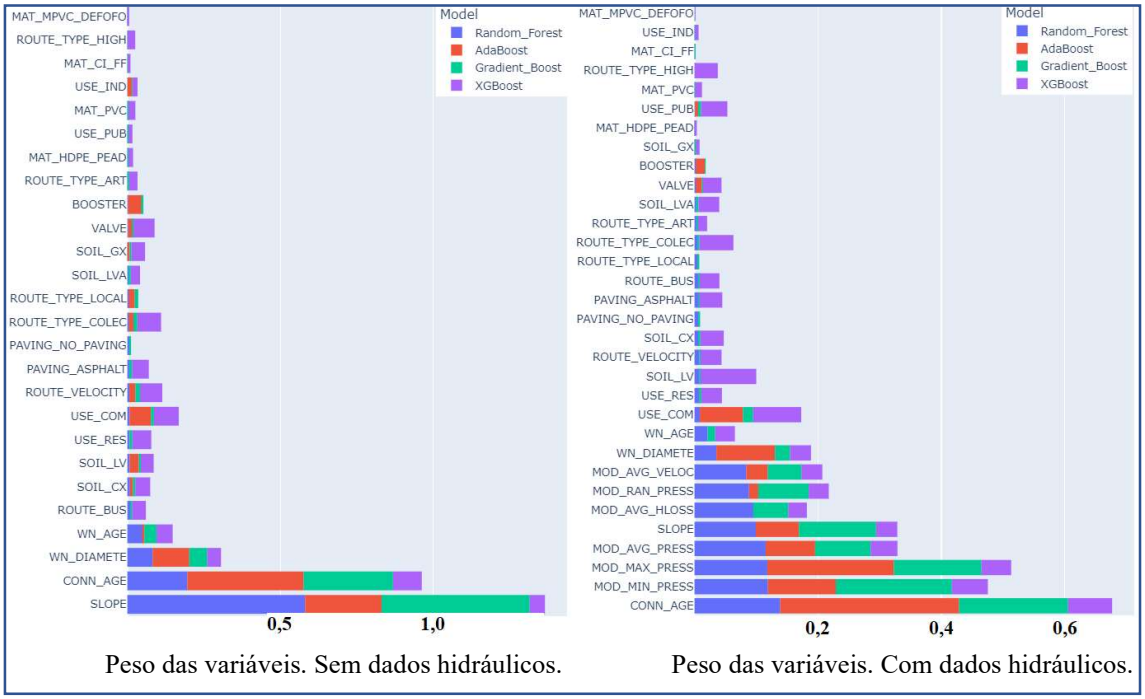
Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.29 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.SSB.001.



Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.30 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.SSB.002.



Peso das variáveis. Sem dados hidráulicos. Peso das variáveis. Com dados hidráulicos.

Figura 5.31 – Variáveis preditoras ao vazamento visível na área de atendimento do RAP.VCP.001.

Tabela 5.8 – Soma do peso das variáveis preditoras segundo os resultados obtidos pelos modelos *Random Forest*, *AdaBoost*, *Gradient Boosting* e *XGBoost*, quanto ao cenário sem a inclusão de dados provenientes de simulação hidráulica das redes de distribuição.

Base de Dados sem variáveis hidráulicas			
Origem da variável preditora	Variável	Descrição	Soma da relevância nos modelos *Random Forest*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*
Aspectos ambientais	SLOPE	Declividade do terreno (%) sob o tubo que provê derivação à ligação	15,544033
Aspectos físicos	CONN_AGE	Idade da conexão/ramal de ligação (ano)	14,783160
Aspectos físicos	WN_AGE	Idade do tubo (ano) que provê derivação à ligação	4,195908
Aspectos físicos	WN_DIAMETE	Diâmetro do tubo (mm) que provê derivação à ligação	3,429118
Aspectos comerciais	USE_RES	Cliente com uso residencial	2,257586
Aspectos operacionais	VALVE	Ligação atendida por VRP	1,424879
Aspectos ambientais	ROUTE_VELOCITY	Velocidade máxima de trânsito na via (km/h)	1,164982
Aspectos comerciais	USE_COM	Cliente com uso comercial	1,158637
Aspectos ambientais	ROUTE_TYPE_COLEC	Ligação em frente a via coletora	1,156727
Aspectos ambientais	SOIL_CX	Ligação sob cambissolo háplico	1,084881
Aspectos operacionais	BOOSTER	Ligação atendida por EBO	1,020837
Aspectos ambientais	SOIL_LV	Ligação sob latossolo vermelho	0,969034
Aspectos ambientais	SOIL_LVA	Ligação sob latossolo vermelho amarelo	0,871167
Aspectos físicos	MAT_PVC	O material do tubo que provê derivação à ligação é PVC	0,835804
Aspectos ambientais	ROUTE_BUS	Ligação em frente a via de circulação de ônibus	0,823182
Aspectos ambientais	PAVING_ASPHALT	Ligação com asfalto à porta	0,803959
Aspectos físicos	MAT_HDPE_PEAD	O material do tubo que provê derivação à ligação é PEAD	0,744107
Aspectos físicos	MAT_CI_FF	O material do tubo que provê derivação à ligação é Ferro Fundido	0,728849
Aspectos ambientais	SOIL_GX	Ligação sob gleissolo háplico	0,578396
Aspectos ambientais	ROUTE_TYPE_ART	Ligação em frente a via arterial	0,510488
Aspectos comerciais	USE_IND	Cliente com uso industrial	0,435289
Aspectos comerciais	USE_PUB	Cliente com uso público	0,388982
Aspectos ambientais	ROUTE_TYPE_HIGH	Ligação em frente a via expressa/rodovia	0,230797
Aspectos físicos	MAT_MPVC_DEFOFO	O material do tubo que provê derivação à ligação é DEFOFO	0,210336
Aspectos ambientais	PAVING_NO_PAVING	Ligação sem asfalto à porta	0,202407
Aspectos ambientais	ROUTE_TYPE_LOCAL	Ligação em frente a via local	0,192509
Aspectos ambientais	SOIL_NV	Ligação sob nitossolo vermelho	0,114940
Aspectos ambientais	ROUTE_TYPE_FAST	Ligação em frente a via rápida	0,079362
Aspectos ambientais	SOIL_RQ	Ligação sob neossolo quartzarênico	0,059644
Aspectos ambientais	SOIL_FF	Ligação sob plintossolo pétrico	0,000000

Tabela 5.9 – Soma do peso das variáveis preditoras segundo os resultados obtidos pelos modelos *Random Forest*, *AdaBoost*, *Gradient Boosting* e *XGBoost*, quanto ao cenário com dados provenientes de simulação hidráulica das redes de distribuição inclusos.

Base de Dados com variáveis hidráulicas			
Origem da variável preditora	Variável	Descrição	Soma da relevância nos modelos *Random Forest*, *AdaBoost*, *Gradient Boosting*, e *XGBoost*
Aspectos físicos	CONN_AGE	Idade da conexão/ramal de ligação (ano)	10,364394
Aspectos operacionais	MOD_MAX_PRESS	Pressão máxima no tubo (mca) que provê derivação à ligação	5,637474
Aspectos operacionais	MOD_RAN_PRESS	Variação de pressão no tubo (mca) que provê derivação à ligação	4,707801
Aspectos operacionais	MOD_AVG_PRESS	Pressão média no tubo (mca) que provê derivação à ligação	4,372775
Aspectos operacionais	MOD_MIN_PRESS	Pressão mínima no tubo (mca) que provê derivação à ligação	4,306769
Aspectos ambientais	SLOPE	Declividade do terreno (%) sob o tubo que provê derivação à ligação	4,219967
Aspectos operacionais	MOD_AVG_HLOSS	Perda de carga média no tubo (m/km) que provê derivação à ligação	3,287524
Aspectos operacionais	MOD_AVG_VELOC	Velocidade média no tubo (m/s) que provê derivação à ligação	2,873573
Aspectos físicos	WN_AGE	Idade do tubo (ano) que provê derivação à ligação	1,884616
Aspectos físicos	WN_DIAMETE	Diâmetro do tubo (mm) que provê derivação à ligação	1,612251
Aspectos operacionais	VALVE	Ligação atendida por VRP	1,466443
Aspectos comerciais	USE_RES	Cliente com uso residencial	1,355999
Aspectos ambientais	SOIL_CX	Ligação sob cambissolo háplico	1,097948
Aspectos comerciais	USE_COM	Cliente com uso comercial	0,995256
Aspectos físicos	MAT_CI_FF	O material do tubo que provê derivação à ligação é Ferro Fundido	0,874464
Aspectos ambientais	ROUTE_VELOCITY	Velocidade máxima de trânsito na via (km/h)	0,862147
Aspectos ambientais	ROUTE_TYPE_COLEC	Ligação em frente a via coletora	0,761738
Aspectos ambientais	ROUTE_BUS	Ligação em frente a via de circulação de ônibus	0,724297
Aspectos ambientais	SOIL_LV	Ligação sob latossolo vermelho	0,651905
Aspectos físicos	MAT_PVC	O material do tubo que provê derivação à ligação é PVC	0,531433
Aspectos ambientais	ROUTE_TYPE_ART	Ligação em frente a via arterial	0,482595
Aspectos físicos	MAT_HDPE_PEAD	O material do tubo que provê derivação à ligação é PEAD	0,455619
Aspectos ambientais	PAVING_ASPHALT	Ligação com asfalto à porta	0,414959
Aspectos ambientais	SOIL_LVA	Ligação sob latossolo vermelho amarelo	0,414885
Aspectos operacionais	BOOSTER	Ligação atendida por EBO	0,375748

Continuação da Tabela 5.10 – Soma do peso das variáveis preditoras segundo os resultados obtidos pelos modelos *Random Forest*, *AdaBoost*, *Gradient Boosting* e *XGBoost*, quanto ao cenário com dados provenientes de simulação hidráulica das redes de distribuição inclusos.

Aspectos comerciais	USE_PUB	Cliente com uso público	0,283497
Aspectos comerciais	USE_IND	Cliente com uso industrial	0,216660
Aspectos ambientais	SOIL_GX	Ligação sob gleissolo háplico	0,157369
Aspectos ambientais	ROUTE_TYPE_LOCAL	Ligação em frente a via local	0,123920
Aspectos ambientais	ROUTE_TYPE_HIGH	Ligação em frente a via expressa/rodovia	0,108692
Aspectos físicos	MAT_MPVC_DEFOFO	O material do tubo que provê derivação à ligação é DEFOFO	0,103551
Aspectos ambientais	PAVING_NO_PAVING	Ligação sem asfalto à porta	0,086007
Aspectos ambientais	ROUTE_TYPE_FAST	Ligação em frente a via rápida	0,080489
Aspectos ambientais	SOIL_NV	Ligação sob nitossolo vermelho	0,063598
Aspectos ambientais	SOIL_RQ	Ligação sob neossolo quartzarênico	0,043632
Aspectos ambientais	SOIL_FF	Ligação sob plintossolo pétrico	0,000000

6. 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, apresentando resultados relevantes para contribuição à tomada de decisão sobre o planejamento de investimentos e a gestão da infraestrutura (Achim *et al*. (2007), Yamijala *et al*. (2009), Jafar *et al*. (2010), Christodoulou (2011), Asnaashari *et al*. (2013), Francis *et al*. (2014), Shirzad *et al*. (2014), Harvey *et al*. (2014), Kabir *et al*. (2015), Aydogdu e Firay (2015), Demissie *et al*. (2017), Farmani *et al*. (2017), Parvizsedghy *et al*. (2017), Kaushik et al. (2017), Sattar et al. (2017), Gómez-Martínez *et al*. (2017), Snider e McBean (2018), Winkler *et al*. (2018), Kumar *et al*. (2018) , Alizadeh *et al*. (2019), Chen *et al*. (2019), Kerwin *et al*. (2019), Wols *et al*. (2019), Robles-Velasco *et al*. (2020), Snider e McBean (2020), Almheiri *et al*. (2020), Giraldo-González e Rodrígues (2020), Snider e McBean (2020b).

A revisão bibliográfica mostrou que há relevância e acurácia superior quando da aplicação de modelos de Aprendizado de Máquina a este tipo de investigação em comparação a métodos estatísticos. Portanto, a presente pesquisa teve como objetivo a aplicação de modelos de Aprendizado de Máquina quanto à falha por vazamento visível em ramais de ligação de água ao ponto de hidrometração de 348.736 ligações no Distrito Federal/Brasil.

A Base de Dados compilada para o estudo foi processada por meio de 12 modelos de Aprendizado de Máquina (*Linear Svm*, *Radial Svm*, *Logistic Regression*, *KNN*, *Decision Tree*, *Naive Bayes*, *Random Forest*, *bagged KNN*, *bagged Decision Tree*, *AdaBoost*, *Gradient Boosting*, e *XGBoost)*. Os resultados obtidos indicaram o modelo *Adaboost* como melhor método para classificação dos dados, performando acurácia de 59,23% utilizando todas as variáveis disponíveis na Base de Dados (aspectos físicos, operacionais, ambientais e comerciais), incluindo aquelas operacionais e provenientes da aplicação de simulação hidráulica. Para ganho final e adicional de performance, o *AdaBoost* foi submetido à hiper-parametrização, o que resultou em acurácia final de 59,70%.

A melhor acurácia observada sem a inclusão das variáveis obtidas por meio de simulação hidráulica (pressões operacionais máximas, mínimas, média e range) foi de 58,51%, também por meio do *AdaBoost* após hiper-parametrização. Portanto, o acréscimo geral de performance proporcionado pelas variáveis hidráulicas foi de 2,03%. O ganho de acurácia corrobora com as investigações realizadas por Robles-Velasco *et al*. (2020) e Snider & McBean (2020b), constatando que os dados provenientes dos modelos hidráulicos também agregam valor a este tipo de investigação quanto à predição de falha por vazamentos.

Em termos de performance, a acurácia dos modelos investigados, em termos gerais, proporcionou o seguinte ranqueamento dos modelos: *AdaBoost*, *XGBoost*, *Gradient Boosting*, *Radial Svm*, *Linear Svm*, *Logistic Regression*, *Random Forest*, *bagged Decision Tree*, *KNN*, *bagged KNN*, *Decision Tree*, e *Naive Bayes*. Assim, as melhores acurácias constatadas foram proporcionadas por meio de *Ensemble Learning Models* do tipo *boosting*.

A Base de Dados permitiu a aplicação dos modelos por áreas de influência da rede de distribuição frente aos reservatórios apoiados dos sistemas de abastecimento de água. As conclusões prestadas acima tangem a média do desempenho dos modelos para os quatorze reservatórios utilizados para a investigação. Avaliando-se a performance por reservatório e, em cinco áreas de atendimento, identificou-se casos em que a acurácia foi superior a 60%. *AdaBoost* após hiper-parametrização obteve desempenho acima de 60% aplicado às áreas de atendimento de cinco reservatórios (RAP.VCP.001, RAP.SSB.002, RAP.SSB.001, RAP.LSL.002, e RAP.PPL.001); *AdaBoost*, quatro reservatórios (RAP.VCP.001, RAP.SSB.002, RAP.SSB.001, e RAP.LSL.002); *Linear* e *Radial SVM*, três reservatórios (RAP.VCP.001, RAP.SSB.002, e RAP.SSB.001); *XGBoost*, dois reservatórios (RAP.VCP.001 e RAP.SSB.002); *Logistic Regression*, dois reservatórios (RAP.SSB.002 e RAP.SSB.001); e, *Gradient Boosting*, um reservatório (RAP.VCP.001). Avaliando-se área de atendimento por área de atendimento, constatou-se acurácias máximas da ordem de 63,6%. Ainda em relação ao desmembramento da Base de Dados por área de atendimento de reservatório apoiado, embora o ganho geral de performance tenha sido da ordem de 2% de melhoria com o uso de dados provenientes de simulação hidráulica, o incremento de acurácia chegou a obter 4% no caso do RAP.SAM.002. Portanto, o processamento de dados por meio de áreas menores, no caso estudado por

área de atendimento de reservatório apoiado, permite melhores acurácias e maior assertividade à aplicação de modelos de Aprendizado de Máquina à falha por vazamento.

Por meio da avaliação de resultados organizados por Matriz de Confusão, constatou-se precisão, quanto à classificação de registros verdadeiros (*ramais que apresentaram vazamento visível reparado*), de 66,01% para o modelo *AdaBoost* após hiper-parametrização.

Quanto às variáveis preditoras, conclui-se que sem dados hidráulicos a idade da conexão e a declividade do terreno sob o tubo da rede de distribuição que provê derivação à ligação de água são as variáveis com maior relevância. No que tange a aplicação considerando variáveis hidráulicas, há significativa predição por meio de elevadas pressões operacionais e idade de rede e ligação de água. Tais considerações reiteram que a redução de pressão, um dos métodos para redução de perdas, mitiga as ocorrências por falha estrutural dos tubos. As demais variáveis operacionais, físicas, ambientais e comerciais compõem o pano de fundo sobre a classificação dos modelos. Observou-se que a análise regionalizada, por reservatório de distribuição, permite que a relevância de diferentes materiais e solos podem agregar acurácia ao processo.

Uma acurácia de 50% indica que não há diferença quanto aos registros de uma variável-alvo, no caso, se houve ou não vazamento visível reparado em uma ligação de água. Em geral, patamares entre 70 e 80% indicam boa acurácia para modelos preditivos, entre 80 e 90%, o desempenho é excelente. Embora a acurácia final não tenha um patamar de excelência, considera-se a aplicação bem-sucedida. Os resultados prestados por meio da classificação por Aprendizado de Máquina podem subsidiar ações para controle das perdas de água por meio do direcionamento de ligações e áreas para intervenções mais assertivas para reabilitação da infraestrutura por substituição de redes e ligações, ou para a realização de pesquisa de vazamento em ramais com maior susceptibilidade ao vazamento. Recomenda-se a aplicação dos métodos de Aprendizado de Máquina avaliados nesta pesquisa em conjunto com outras técnicas que podem direcionar intervenções na rede de distribuição. Por exemplo, o uso de informações provenientes de pesquisa de vazamento por radar indicam áreas com vazamento monitorado remotamente. Tais áreas englobam várias ligações, trata-se da identificação de áreas com vazamento, demandando uma investigação em campo para a localização exata e correção do

vazamento. Assim, as ligações identificadas por meio de Aprendizado de Máquina como mais suscetíveis ao vazamento podem ser as primeiras ligações a serem pesquisadas em campo para a localização do vazamento, o que pode proporcionar acurácia e produtividade às pesquisas, economizando recursos. A instrumentação em redes de distribuição de água por meio de sensores de ruído, de pressão, acelerômetros, ou outros, também constituí hipótese de atuação em conjunto das tecnologias, uma vez que também ajudam a direcionar as ações de investigação em campo, cobrindo áreas, em que as ligações com maior susceptibilidade podem ser pesquisadas primeiro.

No contexto de aplicações no controle de ativo de vazamentos, sugere-se a aplicação dos resultados obtidos nas atividades de pesquisa de vazamentos. O mapeamento das ligações com maior susceptibilidade para ocorrências de vazamento pode direcionar as ações de substituição e renovação da infraestrutura, além de orientar ações para a pesquisa de vazamentos. Neste contexto, um dos componentes da Matriz de Confusão é o Falso Positivo que, nesta pesquisa, traduz-se em ligações classificadas como [1], *ramais que apresentaram vazamento visível reparado*, pelo modelo de Aprendizado de Máquina, mas não houve vazamento visível reparado no período que compõe a Base de Dados. Ou seja, são ligações que possuem as mesmas características daquelas onde houve vazamento visível. Assim, considerando que ocorrem vazamentos visíveis e não visíveis na infraestrutura de sistemas de abastecimento de água, uma hipótese proposta para investigação é que tais ligações possam ter mais susceptibilidade a vazamentos que não afloraram. A identificação das ligações nesta situação pode servir de orientação para a pesquisa de vazamentos, com eventual potencial de direcionamento dos esforços envidados em campo para localização de vazamentos, o que se traduz em possível ganho de performance neste processo.

A compreensão de fatores que determinam a ocorrência de vazamentos em redes de distribuição de água é particularmente crítica para concessionárias por causa de restrições financeiras, aspectos regulatórios e questões ambientais. Como ferramenta para contribuição ao processo decisório sobre gestão de perdas de água, foi apresentada e aplicada uma abordagem que combina métodos de Aprendizado de Máquina e resultados obtidos por simulação hidráulica de redes de distribuição para a predição de vazamentos em ramais de ligação às unidades consumidoras de água. Esta abordagem provê a quantificação e compreensão de como tal elemento de sistemas de abastecimento são

afetados por meio das variáveis que podem induzir ao processo de deterioração da infraestrutura.

À medida que mais dados sejam coletados e sistematizados em Banco de Dados, há a tendências de que tais modelos possam performar maior acurácia à predição. No presente estudo de caso, há pouca informação sobre o ramal de ligação propriamente dito, buscou-se utilizar o material e diâmetro da tubulação que provê derivação ao ramal como meio de considerar características do colar de tomada. Recomenda-se que a instalação de ramais de ligação seja acompanhada de fotos e dados sistematizados em banco sobre as condições de interligação e assentamento do ramal, seu caminhamento e sua chegada ao ponto de instalação do hidrômetro. Maior disponibilidade de dados certamente proverá maior caracterização da infraestrutura e tais condições poderão compor estudos que subsidiem o processo decisório.

REFERÊNCIAS BIBLIOGRÁFICAS

AESBE. (2015). *Série Balanço Hídrico, Guias Práticos,* **vols. 1 a 6**, Associação Brasileira das Empresas Estaduais de Saneamento, Brasília, Brasil.

Achim, D.; Ghotb, F.; McManus, K. J. (2007). "Prediction of Water Pipe Asset Life Using Neural Networks." In: *Journal of Infrastructure Systems*, 13(1), 26-30.

Alegre, H.; Hirner, W.; Baptista, J. M.; Parena, R. (2000). *Performance Indicators for Water Supply Systems. Series: Manual of Best Practice.* IWA Publishing, Londres, Reino Unido.

Alegre, H.; Covas, D.I.C.; Coelho, S.T.; Almeida, M.C.; Cardoso, M.A. (2012). "An integrated approach for infrastructure asset management of urban water systems." In: *Water Asset Management International*, 8(2), 10-14.

Alegre, H.; Baptista, J. M.; Cabrera Jr, E. C.; Cubillo, P. D.; Hirner, W.; Wolf, M.; Parena, R. (2017). *Performance Indicators for Water Supply Systems. Series: Manuals of Best Practice*, 3rd Edition, IWA Publishing.

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

Almeida, M.C., Leitão, J.P., Coelho, S.T. (2011). "Risk management in urban water infrastructures: application to water and wastewater systems." In Almeida, B. (ed.), *Water management, uncertainty and risks: operational conceptualization.* Esfera do Caos, Lisbon, Portugal.

Almheiri, Z.; Meguid, M.; Zayed, T. (2020). "Intelligent Approaches for Predicting Failure of Water Mains". In: *J. Pipeline Syst. Eng. Pract.*, 11(4), 04020044.

Andrews, S.; Tsochantaridis, I.; Hofmann, T. (2003) "Support vector machines for multiple-instance learning." In: *Adv. Neural Inf. Process. Syst.*, 15, 561-568.

Asadzadeh, M.; Tolson, B.A.; McKillop, R. (2011). "A Two Stage Optimization Approach for Calibrating Water Distribution Systems." In: *Water Distribution Systems Analysis*; American Society of Civil Engineers, Tucson, AZ, USA, 1682-1694.

Asnaashari, A.; McBean, E. A.; Gharabaghi, B.; Tutt, D. (2013). "Forecasting Watermain Failure Using Artificial Neural Network Modelling." In: *Canadian Water Resources Journal*, 38(1), 24-33.

ABNT, Associação Brasileira de Normas Técnicas. 2017. *NBR 12218: Projeto de rede de distribuição de água para abastecimento público.* Rio de Janeiro, Brasil.

Aydogdu, M.; Firat, M. (2015). "Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods". In: *Water Resour. Manage*, 29, 1575-1590.

AWWA, American Water Works Association. (2009). *Loss Control Programs-Manual of Water Supply Practices*, M36, Denver, USA.

AWWA, American Water Works Association. (2016). *Water Audits and Loss Control Programs*, M36, 4th edition, USA.

Berardi, L.; Kapelan, O.; Giustolisi, O.; Savic, D. (2008). "Development of Pipe Deterioration Models for Water Distribution Systems Using EPR." In: *Journal of Hydroinformatics*, 10 (2), 113-126.

Bergstra, J.; Bengio, Y. (2012). "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research*, 13, 281-305.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Bhave, P.R. (1988). "Calibrating Water Distribution Network Models." In: *J. Environ. Eng.*, 114, 120-136.

Breiman, L. (2001). "Random forests." In: *Machine Learning*, 45(1), 5-32.

Campesato, O. (2020a). *Artificial Intelligence: Machine Learning and Deep Learning*, Stylus Publishing, LLC.

Campesato, O. (2020b). *Python 3 for Machine Learning*, Stylus Publishing, LLC.

Charalambous, B.; Fouferas, D.; Petroulias, N. (2014). "Leak detection and water loss management." In: *Water Utility Journal*, 8, 25-30.

Chen, T.; Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*. ACM: New York, NY, USA, 785-794.

Chen, Y. T.; Beekman, J. A.; Guikema, S. D.; Shashaani, S. (2019). "Statistical Modeling in Absence of System Specific Data: Exploratory Empirical Analysis for Prediction of Water Main Breaks". In: *ASCE J. Infrastruct. Syst.*, 25(2), 04019009.

Christodoulou, S. E. (2010). "Water network assessment and reliability analysis by use of survival analysis." In: *Water Resour. Manage.* 25(4), 1229-1238.

Claesen, M.; De Moor, B. (2015). *Hyperparameter Search in Machine Learning*. arXiv:1502.02127.

Collins, M.; Cooper, L.; Helgason, R.; Kennington, J.; Leblanc, L. (1978). “Solving the Pipe Network Analysis Problem Using Optimization Techniques.” In: *Manag. Sci.*, 24, 747-760.

Cortes, C.; Vapnik, V. N. (1995). “Support-vector networks.” In: *Machine Learning*, 20(3), 273-297.

Covelli, C.; Cozzolino, L.; Cimorelli, L.; Della Morte, R.; Pianese, D. (2015). “A model to simulate leakage through joints in water distribution systems.” In: *Water Supply*, 15, 852-863.

Covelli, C.; Cimorelli, L.; Cozzolino, L.; Della Morte, R.; Pianese, D. (2016). “Reduction in water losses in water distribution systems using pressure reduction valves.” In: *Water Supply*, 16, 1033-1045.

Dangeti, P. (2017). *Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R*. 1st ed., Packt Publishing Ltd.

Darlington, R.B. (1990). *Regression and Linear Models*. Columbus, OH: McGraw-Hill Publishing Company.

Davies, E. R. (2018). *Computer Vision: Principles, Algorithms, Applications, Learning*. 5th ed., Academic Press: Elsevier.

Davis, J.; Goadrich, M. (2006). “The relationship between precision-recall and roc curves.” In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM, New York, NY, USA, 233-240.

Demissie, G.; Tesfamariam, S.; Sadiq, R. (2017). “Prediction of pipe failure by considering time-dependent factors: Dynamic Bayesian belief network model.” In: *ASCE-ASME J. Risk Uncertainty Eng. Syst. Part A Civ. Eng.*, 3(4), 04017017.

Di Nardo, A.; Di Natale, M.; Gisonni, C.; Iervolino, M. (2014). “A genetic algorithm for demand pattern and leakage 5. estimation in a water distribution network.” In: *J. Water Supply Res. Technol.*, 64, 35-46.

Do, N.C.; Simpson, A.R.; Deuerlein, J.W.; Piller, O. (2016). “Calibration of Water Demand Multipliers in Water Distribution Systems Using Genetic Algorithms.” In: *J. Water Resour. Plan. Manag.*, 142, 04016044.

Do, N.C.; Simpson, A.R.; Deuerlein, J.; Piller, O. (2017). “Particle Filter–Based Model for Online Estimation of Demand Multipliers in Water Distribution Systems under Uncertainty.” In: *J. Water Resour. Plan. Manag.*, 143, 04017065.

Drucker, H.; Burges, C.; Kaufman, L.; Smola, A.; Vapnik, V. (1997); "Support Vector Regression Machines". In: *Advances in Neural Information Processing Systems*, 9, 155-161.

European Commission, D. G. f. t. E, (2015). *EU Reference Document Good Practices on Leakage Management WFD CIS WG PoM: Main Report*. Brussels, European Union.

Fanner, P.; Thornton, J. (2005). "The importance of real loss component analysis for determining the correct intervention strategy." In: *Proceedings of IWA Water Loss 2005 Conference*. Halifax, Nova Scotia, Canada.

Farley, M.; Trow, S. (2003). *Losses in Water Distribution Networks*. IWA Publishing, London, UK.

Farley, M.; Wyeth, G.; Ghazali, Z. B. M.; Istandar, A.; Singh, S.; Dijk, N.; Raksakulthai, V.; Kirkwood, E. (2008). *The Manager's Non-revenue Water Handbook: A Guide to Understanding Water Losses*. Ranhill Utilities Berhad and the United States Agency for International Developement, Bangkok, Thailand.

Farmani, R.; Kakoudakis, K.; Behzadian, K.; Butler, D. (2017). "Pipe failure prediction in water distribution systems considering static and dynamic factors." In: *Procedia Eng.*, 186, 117-126.

Fawcett, T. (2006). "An introduction to ROC analysis." In: *Pattern Recognit. Lett.*, 27, 861874.

Francis, R. A.; Guikema, S. D.; Henneman, L. (2014). "Bayesian belief networks for predicting drinking water distribution system pipe breaks." In: *Reliab. Eng. Syst. Saf.*, 130, 1-11.

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine." In: *Annals of Statistics*, 1189-1232.

Giraldo-González, M. M.; Rodrígues, J. P. (2020). "Comparison of Statistical and Machine Learning Models for Pipe Failure Modeling in Water Distribution Networks". In: *MDPI Water Journal*, 12, 1153.

Giustolisi, O; Laucelli, D.; Dragan, A. (2006). "Development of rehabilitation plans for water mains replacement considering risk and cost benefit assessment." In: *J. Civ. Eng. Environ. Syst.*, 23(3), 175-190.

GIZ, Deutsche Gesellschaft für Internationale Zusammenarbeit. (2011). *Guidelines for water loss reduction: a focus on pressure management.* Eschborn, Alemanha.

Gómez-Martínez, P.; Cubillo, F.; Martín-Carrasco, F. J.; Garrote, L. (2017). “Statistical Dependence of Pipe Breaks on Explanatory Variables”. In: *MDPI. Water*, 9, 158.

Gonçalves, E. (1998). *Metodologias para Controle de Perdas em Sistemas de Distribuição de Água - Estudo de Casos da CAESB*, Dissertação de Mestrado, Publicação MTARH.DM - 010A/98, Departamento de Engenharia Civil, Universidade de Brasília, Brasília, DF, 173p.

Gönen, M. (2007). *Analyzing Receiver Operating Characteristic Curves with SAS*. Cary, NC: SAS Institute Inc.

Gul, A.; Perperoglou, A.; Khan, Z. (2018). “Ensemble of a subset of kNN classifiers.” In: *Adv. Data Anal Classif.*, 12, 827-840.

Gouveia, C. G.N; Soares, A. K. (2021). “Contribuições à aplicação de modelos hidráulicos em setorização de redes de distribuição de água em casos reais.” In: *31° Congresso da Associação Brasileira de Engenharia Sanitária e Ambiental.* Curitiba, Paraná, Brasil.

Hand, D. J.; Yu, K. (2001). "Idiot's Bayes - not so stupid after all?". In: *International Statistical Review,* 69(3), 385-399.

Harrison, M.; Petrou, T. (2020). *Pandas 1.x Cookbook, Practical Recipes for Scientific Computing, Time Series Analysis, and Exploratory Data Analysis using Python*, 2nd ed., Packt Publishing, Birmingham, UK.

Harvey, R.; McBean, E. A.; Gharabaghi, B. (2014). “Predicting the timing of water main failure using artificial neural networks.” In: *J. Water Resour. Plan. Manag.*, 140, 425-434.

Hirner, W.; Lambert, A. (2000). *Losses from Water Supply Systems: Standard Terminology and Recommended Performance Measures*. IWA Blue Pages, London, UK.

Hosmer, D.W.; Lemeshow, S.L. (2000). *Applied Logistic Regression*. 2nd ed., Hoboken, NJ: Wiley-Interscience.

Hutton, C.J.; Kapelan, Z.; Vamvakeridou-Lyroudia, L.; Savic, D. (2014). “Dealing with Uncertainty in Water Distribution System Models: A Framework for Real-Time Modeling and Data Assimilation.” In: *J. Water Resour. Plan. Manag.*, 140, 169-183.

InfraGuide (2003). *Deterioration and Inspection of Water Distribution Systems*, Federation of Canadian Municipalities and National Research Council, Canadá.

Itonaga, L. C. H. (2005). *Estudo da Aplicação de Modelos de Redes de Água no Controle de Perdas em Casos Reais*, Dissertação de Mestrado, Publicação PTARH.DM – 80/2005, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 201 p.

Jafar, R.; Shahrour, I.; Juran, I. (2010). "Application of Artificial Neural Networks (ANN) to Model the Failure of Urban Water Mains." In: *Mathematical and Computer Modelling*, 51(9–10), 1170–1180.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. (2013). *An introduction to statistical learning*, **vol. 6**. Springer, Berlin.

Jiang, L.; Cai, Z.; Wang, D. (2014). "Bayesian Citation-KNN with distance weighting." In: *Int. J. Mach. Learn. & Cyber*, 5, 193-199.

Kaddoura, K.; Zayed, T. (2018). "Erosion void condition prediction models for buried linear assets." J. In: *Pipeline Syst. Eng. Pract.*, 10(1), 04018029.

Kabir, G.; Tesfamariam, S; Sadiq, R. (2015). "Bayesian model averaging for the prediction of water main failure for small to large Canadian municipalities." In: *Can. J. Civ. Eng.* 43 (3), 233-240.

Kakoudakis, K.; Behzadian, K.; Farmani, R.; Butler, D. (2017). "Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K-means clustering." In: *Urban Water J.*, 14, 737-742.

Kamel, S.; Meguid, M. A. (2012). "Investigating the effects of local contact loss on the earth pressure distribution on rigid pipes." In: *Geotech. Geol. Eng.*, 31(1), 199-212.

Kang, D.; Lansey, K. (2011). "Demand and Roughness Estimation in Water Distribution Systems." In: *J. Water Resour. Plan. Manag.*, 137, 20-30.

Kaur, H.; Pannu, H. S.; Malhi, A. K. (2019). "A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions". In: *ACM Comput. Surv*. 52, 4, Article 79.

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

Kerwin, S.; Soto, B. G.; Adey, B. T. (2019). "Performance Comparison for Pipe Failure Prediction Using Artificial Neural Networks". In: Caspeele, Taerwe & Frangopol (eds.) *Life-Cycle Analysis and Assessment in Civil Engineering: Towards an Integrated Vision*, Taylor & Francis Group, London.

Kingdom, B.; Liemberger, R.; Marin, P. (2006). “The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries. How the Private Sector Can Help: A Look at Performance-Based Service Contracting.” In: *Water Supply and Sanitation Board Discussion Paper Series*, 8, World Bank, Washington, DC.

Kishawy, H. A.; Gabbar, H. A. (2010). “Review of pipeline integrity management practices.” In: *Int. J. Press. Vessels Pip.*, 87(7), 373-380.

Kotsiantis S. B.; Tsekouras G.E.; Pintelas P.E. (2005). “Bagging Model Trees for Classification Problems.” In: Bozanis P., Houstis E.N. (eds) *Advances in Informatics.* PCI 2005, Lecture Notes in Computer Science, vol 3746. Springer, Berlin Heidelberg.

Kramer, O. (2013). *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Intelligent Systems Reference Library, **vol. 51**, Springer, Berlim.

Kumar, A.; Rizvi, S.; Brooks, B.; Vanderveld, R.; Wilson, K.; Kenney, C.; Edelstein, S. (2018). “Using Machine Learning to Assess the Risk of and Prevent Water Main Breaks.” In: *SIGKDD’18*, 1-9. London, UK.

Kumar, A.; Jain, M. (2020). *Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases*. Apress, Berkeley, CA. ISBN-13 (pbk): 978-1-4842-5939-9.

Lambert, A. (1994). *Managing Leakage: Interpreting measured night flows - Report E*. Water Research Centre Publications.

Lambert, A.; Hirner, W. (2000). *Losses from Water Supply Systems: Standard Terminology and Recommended Performance Measures*. IWA Blue Pages.

Lambert, A.; Brown, T.G.; Takizawa, M; Weimer, D., (1999). “A review of performance indicatiors for real losses from water supply systems.” In: *J. Water Supply Res. Technol.*, 48 (6), 227-237.

Lambert, A.; Charalambous, B.; Fantozzi, M.; Kovac, J.; Rizzo, A.; St John, S. G. (2014). “14 years’ experience of using IWA best practice water balance and water loss performance indicators in Europe.” In: *Proceedings of IWA Specialized Conference*: Water Loss 2014, Vienna.

Laucelli, D.; Rajani, B.; Kleiner, Y.; Giustolisi, O. (2014). “Study on Relationships between Climate-Related Covariates and Pipe Bursts Using Evolutionary-Based Modelling.” In: *Journal of Hydroinformatics*, 16(4), 743.

Laucelli, D.; Berardi, L.; Giustolisi, O.; Vamvakeridou-Lyroudia, L.S.; Kapelan, Z.; Savic, D.; Barbaroand, G. (2011). “Calibration of Water Distribution System Using

Topological Analysis.” In: *Water Distribution Systems Analysis*, American Society of Civil Engineers: Tucson, AZ, USA, 1664-1681.

Letting, L.; Hamam, Y.; Abu-Mahfouz, A.M. (2017). “Estimation of Water Demand in Water Distribution Systems Using Particle Swarm Optimization.” In: *Water 2017*, 9, 593.

Liemberger, R.; Wyatt, A. (2018). “Quantifying the Global Non-revenue Water Problem.” In: *Water Science & Technology Water Supply*, 19(3).

Liu, Z.; Kleiner, Y. (2014). “Computational Intelligence for Urban Infrastructure Condition Assessment: Water transmission and distribution systems.” In: *IEEE Sens. J.*, 14(12), 4122-4133.

Makar, J.; Desnoyers, R.; McDonald, S. (2001). “Failure modes and mechanisms in gray cast iron pipe.” In: *Proc., Underground Infrastructure Research*, 1-10. Kitchener, ON: A.A. Balkema.

Mays, L. (2000). *Water distribution systems handbook*. McGraw Hill Professional.

McCullagh, P.; Nelder, J. (1989). *Generalized linear models*, **vol. 37**, CRC press.

Meirelles, G.; Manzi, D.; Brentan, B.M.; Goulart, T.; Junior, E.L. (2017). “Calibration Model for Water Distribution Network Using Pressures Estimated by Artificial Neural Networks.” In: *Water Resour. Manag.*, 31, 4339-4351.

Mohri, M.; Rostamizadeh, A.; Talwalkar, A. (2018). *Foundations of Machine Learning*, MIT Press.

Mukhiya, S. K.; Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python*. Packt Publishing, Birmingham, UK.

Mutikanga, H. E.; Sharma, S. K.; Vairavamoorthy, K. (2013). “Methods and tools for managing losses in water distribution systems.” In: *J. Water Resour. Plan. Manage.*, 139(2), 166-174.

Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". In: *Journal of Artificial Intelligence Research*, 11, 169-198. doi: 10.1613/jair.614.

Ormsbee, L. E.; Wood, D. J. (1986). “Explicit Pipe Network Calibration.” In: *J. Water Resour. Plan. Manag.*, 112, 166-182.

Parvizsedghy, L., I. Gkountis, A. Senouci, T. Zayed, M. Alsharqawi, H. El Chanati, M. El-Abbasy, and F. Mosleh. (2017). “Deterioration assessment models for water pipelines.” In: *Int. J. Civ. Environ. Eng.*, 11(7), 1013-1022.

Polikar, R. (2006). "Ensemble based systems in decision making". In: *IEEE Circuits and Systems Magazine*, 6(3), 21-45.

Powers, D. M. (2011). *Evaluation: from precision, recall and f-measure to roc, informedness markedness and correlation*.

Puust, R.; Kapelan, Z.; Savic, D.; Koppel, T. (2010). “A review of methods for leakage management in pipe networks.” In: *Urban Water J.*, 7, 25-45.

Quinlan, J. R. (1986). "Induction of decision trees". In: *Machine Learning*, 1(1), 81-106.

Rajani, B.; Kleiner, Y. (2001). “Comprehensive review of structural deterioration of water mains: Physically based models.” In: *Urban Water*, 3(3), 151-164.

Rajani, B.; Kleiner, Y. (2001b). “Comprehensive review of structural deterioration of water mains: Statistical models.” In: *Urban Water*, 3, 131-150.

Rajani, B., Y. Kleiner (2002). "Towards Pro-active Rehabilitation Planning of Water Supply Systems." In: *International Conference on Computer Rehabilitation of Water Networks-CARE-W*, Dresden, Germany.

Robles-Velasco, A.; Cortés, P.; Muñuzuri, J.; Onieva, L. (2020). “Prediction of pipe failures in water supply networks using logistic regression and support vector classification.” In: *Reliability Engineering & System Safety*, 196, 106754.

Rokach, L.; Maimon, O. (2008). *Data mining with decision trees: theory and applications*, World Scientific Pub Co Inc.

Rokach, L. (2010). "Ensemble-based classifiers". In: *Artificial Intelligence Review*, 33(1-2), 1-39.

Rokach, L. (2019). *Ensemble Learning, Pattern Classification Using Ensemble Methods. Series in Machine Perception and Artificial* Intelligence, **vol. 85**, World Scientific Publishing Co. Pte. Ltd.

Rossman, L. A. (2000) *EPANET 2: User’s Manual*. Available online: https://nepis.epa.gov/Adobe/PDF/P1007WWU.pdf.

Russell, S.; Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*, 2nd ed., Prentice Hall.

Russel, R. (2018). *Machine Learning: Step-by-Step Guide to Implement Machine Learning Algorithms with Python*, CreateSpace Independent Publishing Platform.

Savic, D.; Giustolisi, O.; Laucelli, D. (2009a). “Asset Deterioration Analysis Using Multi-Utility Data and Multi-Objective Data Mining.” In: *Journal of Hydroinformatics*, 11(3-4), 211-224.

Savic, D.; Kapelan, Z. S.; Jonkergouw, P.M. (2009b). “Quo vadis water distribution model calibration?” In: *Urban Water J.*, 1(6), 3-22.

Sattar, A. M.; Ertuğrul, Ö. F.; Gharabaghi, B.; McBean, E.; Cao, J. (2017). "Extreme learning machine model for water network management." In: *Neural Comput. Appl.* 31(1), 157-169.

Schapire R.E. (2013) "Explaining AdaBoost". In: Schölkopf B., Luo Z., Vovk V. (eds) *Empirical Inference*. Springer, Berlin, Heidelberg.

Scheidegger, A.; Leitão, J. P.; Scholten, L. (2015). "Statistical failure models for water distribution pipes - A review from a unified perspective." In: *Water Res.*, 83, 237-247.

Serranito, F. S.; Donnelly, A. (2015). Active Water Loss Control. EPAL, Empresa Portuguesa das Águas Livres S.A., Lisbon, 95p.

Shirzad, A.; Tabesh, M.; 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." In: *KSCE J. Civ. Eng.*, 18(4), 941-948.

Snider, B.; McBean, E. A. (2018). "Improving Time to Failure Predictions for Water Distribution Systems Using Gradient Boosting Algorithm." In: *WDSA/CCWI Joint Conference Proceedings*, Kingston, Canada.

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

Snider, B.; McBean, E. A. (2020b). "Improving Urban Water Security through Pipe-Break Prediction Models: Machine Learning or Survival Analysis." In: *J. Environ. Eng.*, 146(3), 04019129.

SNIS (2019). *Sistema Nacional de Informações Sobre Saneamento, Diagnóstico dos Serviços de Água e Esgotos*, **Capítulo 8**. Ministério Desenvolvimento Regional.

Soares, A. K. (2003). *Calibração de Modelos de Redes de Distribuição de Água para Abastecimento Considerando Vazamentos e Demandas Dirigidas pela Pressão*, Dissertação de Mestrado, Universidade de São Paulo, Escola de Engenharia de São Carlos, São Carlos, 153 p.

Taal Water District. (2016). *Water service connection*. [online]. Disponível na Internet via: https://taalwd.gov.ph/services/water-service-connection.

Tabesh, M.; Soltani, J.; Farmani, R.; Savic, D. (2009). "Assessing Pipe Failure Rate and Mechanical Reliability of Water Distribution Networks Using Data-Driven Modeling." In: *Journal of Hydroinformatics*, 11(1), 1-17.

Tabesh, M.; Jamasb, M.; Moeini, R. (2011). “Calibration of water distribution hydraulic models: A comparison between pressure dependent and demand driven analyses.” In: *Urban Water J.*, 8, 93-102.

Todini, E.; Pilati, S. (1988). “A gradient algorithm for the analysis of pipe networks.” In: Coulbeck, B.; Orr, C. (eds.) *Computer Applications in Water Supply: Vol. 1 - Systems Analysis and Simulation*, Research Studies Press Ltd., Baldock, UK, 1-20.

Trow, S.; Hall, M. (1994). *Managing Leakage - Setting Economic Leakage Targets Repor C*, Water Research Centre Publications, Londres, Reino Unido, 125 p.

Trow, S.; Tooms, S. (2014). “Pressure Management of Water Distribution Systems: Every Metre Counts.” In: *Conference Water Ideas 2014*, Bologna, Italy.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass: Addison-Wesley Pub.

Unpingco, J. (2016). *Python for Probability, Statistics, and Machine Learning*, Springer.

Walski, T. M. (1983). “Technique for Calibrating Network Models.” In: *J. Water Resour. Plan. Manag.*, 109, 360-372.

Wang, Y.; Zayed, T.; Moselhi, O. (2009). “Prediction models for annual break rates of water mains.” In: *J. Perform. Constr. Facil.*, 23(1), 47-54.

Wilson, D.; Filion, Y.; Moore, I. (2015). “State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains.” In: *Urban Water J.*, 14, 173-184.

Winkler, D.; Haltmeier, M.; Kleidorfer, M.; Rauch, W.; Tscheikner-Gratl, F. (2018). “Pipe Failure Modelling for Water Distribution Networks Using Boosted Decision Trees.” In: *Structure and Infrastructure Engineering*, 14(10), 1402-1411.

Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*, Elsevier, Amsterdam, Netherlands.

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

Wu, Y.; Liu, S. (2017). “A review of data-driven approaches for burst detection in water distribution systems.” In: *Urban Water J.*, 14, 972-983.

Wu, Z.Y.; Walski, T.; Mankowski, R.; Herrin, G.; Gurrieri, R.; Tryby, M. (2002). “Calibrating water distribution model via genetic algorithms.” In: *Proceedings of the 2002 AWWA IMTech Conference*, Kansas, MO, USA, 14-17.

Yamijala, S.; Guikema, S. D.; Brumbelow, K. (2009). “Statistical models for the analysis of water distribution system pipe break data.” In: *Reliab. Eng. Syst. Saf.*, 94(2), 282-293.

Zhang, C.; Ma, Y. (2014). *Ensemble Machine Learning, Methods and Applications*. Springer, London.

Zhou, J.; Chen, F. (2018). *Human and machine learning: Visible, explainable, trustworthy and transparent*, Springer, Cham, Switzerland.

Zhou, X.; Xu, W.; Xin, K.; Yan, H.; Tao, T. (2018). “Self-Adaptive Calibration of Real-Time Demand and Roughness of Water Distribution Systems.” In: *Water Resour. Res.*, 54, 5536-5550.

Zhou, Z.H. (2004). *Multi-instance learning: a survey*. Technical Report, AI Lab, Department of Computer Science and Technology, Nanjing University, Nanjing.