DETECÇÃO DE FRAUDES EMPREGANDO *DATA SCIENCE TRAJECTORY* (DST), *EXTREME GRADIENT BOOSTING* (XGBOOST) E *SHAPLEY ADITIVE EXPLANATIONS* (SHAP): UMA NOVA PROPOSTA.

Gustavo de Souza Groppo

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

Endereço: Rua Maranhão, 987, apto 601 - Funcionários - Belo Horizonte - MG - CEP: 30150-331 - Brasil - Tel: (31) 97195-9955 - e-mail: gustavo.groppo@gmail.com

RESUMO: As perdas advindas do consumo não autorizado de água são um problema importante para o setor de saneamento, visto que parte da água que é produzida pela concessionária não é faturada. É um tema de alta relevância frente aos cenários de escassez hídrica e de altos custos de energia elétrica, impactando na saúde financeira dos prestadores, uma vez que podem representar desperdício de recursos naturais, operacionais e financeiros. Propõe-se uma metodologia de caracterização espacial das fraudes que leva em consideração a logística comercial de leitura dos hidrômetros. Para tal emprega-se a trajetória da ciência de dados, juntamente com o algoritmo XGBoost e o método SHAP.

PALAVRAS-CHAVE: Perdas; Meio Ambiente; Sustentabilidade.

1. INTRODUÇÃO

Apesar de todas as infraestruturas hídricas urbanas existentes, muitas cidades estão sob estresse hídrico. Uma grande proporção da população mundial tem sido afetada [Vörösmarty *et al.*, 2000, Maddocks *et al.*, 2015], impondo o risco de escassez para aproximadamente um quarto da população mundial [Mcdonald *et al.* 2014]. Esta escassez passou a ser um dos fatores mais críticos no processo de gestão de algumas empresas diante de fatores como mudanças climáticas, aumento das taxas de urbanização e crescimento da renda e da produção [Molinos-Senante *et al.*, 2016] e das perdas dos sistemas de abastecimento de água (SAA) [Liemberger e Wyatt, 2019].

Neste ambiente de escassez hídrica, as perdas provocam um grande impacto no ambiente em função do elevado volume de água extraído dos mananciais, repercutindo significativamente na sustentabilidade de abastecimento. Estima-se que o nível médio global de perdas foi de aproximadamente 35% da água produzida que responde por 48,6 bilhões de metros cúbicos por ano (m^3/ano), refletindo diretamente no balanço financeiro das concessionárias [Kingdom *et al.*, 2006; Farley *et al*., 2008]. Contudo, em estudo recente [Liemberger e Wyatt, 2019], o volume global de água não faturada (NRW) foi de 346 milhões de m^3/dia ou 126 bilhões de m^3/ano, totalizando US$ 39 bilhões por ano de perda em receita. No contexto da América Latina e região do Caribe, o Brasil é o país com maior nível de NRW [Liemberger e Wyatt, 2019].

As perdas advindas do consumo não autorizado de água, chamadas de perdas não físicas ou comerciais, são um problema importante para o setor de saneamento, visto que parte da água que é produzida pela concessionária não é faturada. É um tema de alta relevância frente aos cenários de escassez hídrica e de altos custos de energia elétrica, além da sua relação direta com a saúde financeira dos prestadores, uma vez que podem representar desperdício de recursos naturais, operacionais e financeiros [Brasil 2020]. Estas perdas incluem a imprecisão da micromedição (hidrômetros inoperantes, com submedição, erros de leitura, fraudes, equívocos na calibração do hidrômetro), ligações clandestinas, *by pass* irregulares nos ramais das ligações, falhas no cadastro comercial e outras situações [Seago e McKenzie, 2007; Fanner *et al.*, 2007; Fettermann *et al.*, 2015; Monedero *et al.*, 2016; Alegre *et al.*, 2016]. Cabe salientar que estas perdas não representam necessariamente um volume de água que não foi utilizada, assim como nas perdas reais. Entretanto, esses tipos de perdas além de serem difíceis de detectar também afetam diretamente o equilíbrio entre o volume de água que entra e sai de um SAA [Lambert e Hirner, 2010].

Os fatores que influenciam esse tipo de perda podem estar associados a questões sociais, culturais, financeiras e políticas da região de abastecimento [Farley e Liemberger, 2005]. A população de áreas vulneráveis acaba sendo induzida a realizarem conexões clandestinas [Bharti *et al.*, 2020], uma das principais causas dos elevados níveis de perda aparente de água, podendo atingir, em algumas regiões, um percentual de ligações clandestinas na ordem 26% [Seago e McKenzie, 2007]. Parte dessa perda acontece pela combinação de falta de instrução e conhecimento dos usuários e pela ausência de políticas de gestão [Gonzáles-Gomes *et al.*, 2012]. Em locais onde a população participa ativamente da manutenção de estratégias de gestão, denunciando ligações clandestinas, geralmente obtêm relativo êxito na redução da água não faturada [Lai *et al.*, 2016]. Políticas semelhantes para redução de perdas aparentes foram adotadas no Camboja [Biswas e Tortajada, 2010].

Nessa perspectiva, a comunidade científica vem estudando alternativas que impactem na redução do NRW. A maioria dos estudos estão relacionadas as perdas técnicas. Esta perda está ligada com os problemas no sistema de distribuição (vazamentos na adução e distribuição, vazamentos e transbordamentos em tanques de armazenamento e vazamentos em ramais de ligações e conexões) [Mounce *et al.*, 2011; Mashford *et al.*, 2012; Srirangarajan *et al.*, 2013; Meseguer *et al.*, 2014; Jung *et al.*, 2015; Jang *et al.*, 2018]. No tocante as perdas aparentes (consumo não autorizado e imprecisões nos medidores), poucos são os estudos acerca do tema [Mutikanga *et al.*, 2011; Humaid e Barhoum, 2013; Fettermann *et al.*, 2015; Garcia *et al.*, 2015; Monedero *et al.*, 2016; Morote e Hernández-Hernández, 2018; Al-Radaideh e Al-Zoubi, 2018; Trancoso *et al.*, 2020; Nofal *et al.*, 2021; Kainz *et al.*, 2021].

Em alguns problemas específicos, a localização espacial dos fenômenos é muito importante e em alguns casos imprescindível à completa compreensão do problema, permitindo o conhecimento a respeito das características das sub-regiões nas quais se localizam as perdas comerciais. Tal conhecimento é determinante para maior eficiência na prevenção e no combate a tais perdas. Logo, há fortes indícios de que o lugar onde ocorrem as perdas comerciais é importante e deve compor a análise do problema [Faria *et al.*, 2014].

Dado o exposto, propõe-se a caracterização espacial das fraudes, empregando a logística comercial de leitura dos hidrômetros, e o desenvolvimento de um modelo de predição de ocorrência de consumo não autorizado de água empregando a trajetória de ciência de dados (DST), juntamente com o algoritmo *e**X**treme **G**radient **B**oosting* (XGBoost) [Chen e Guestrin, 2016] e o método ***SH**apley **A**ditive ex**P**lanations* (SHAP) [Lundberg e Lee, 2017].

2. ESTUDOS CORRELATOS

As concessionárias de abastecimento de água têm buscado cada vez mais soluções para reduzir o desperdício de água, principalmente em um ambiente de escassez hídrica. Muitos esforços têm sido feitos com o objetivo de promover uma melhor gestão deste recurso. Contudo, a literatura disponível, relacionado à detecção de fraudes no consumo de água, é limitada quando se comparadas com outros setores, como o setor elétrico e financeiro por exemplo.

Fettermann *et al.* (2015) propôs uma sistemática para direcionar a inspeção dos possíveis locais com fraude antes realizado aleatoriamente pela companhia. Os resultados obtidos apontam que a aplicação da sistemática proporciona uma taxa de detecção de 98% dos casos de fraude no consumo de água, se mostrando uma alternativa capaz de direcionar os locais a serem inspecionados quanto à existência de irregularidades, apesar da taxa de erro obtido pelo método de Z-score modificado ser considerada ainda alta.

Monedero *et al.* (2015) desenvolveram uma metodologia constituída por um conjunto de três algoritmos para a detecção de adulteração de medidores na Empresa Metropolitana de Abastecimiento e Saneamiento de Águas de Sevilla (EMASESA). Os algoritmos foram gerados e programados após um processo de mineração de dados no banco de dados da empresa detectando três tipos de padrões de consumo (quedas progressivas, quedas repentinas e consumo anormalmente baixo). Esta metodologia foi testada com inspeções *in loco* a clientes de uma aldeia da província de Sevilha. Após a realização das inspeções pela concessionária, os fiscais confirmaram um bom índice de sucesso, tendo em vista que a detecção desse tipo de fraude é muito difícil por se tratar de uma técnica não invasiva. Assim, como em Mondedro *et al.* (2015), Monedero *et al.* (2016) desenvolveram uma metodologia que possibilita detectar manipulações nos medidores incluindo informações geográficas dos pontos de abastecimento do cliente visando melhorar os resultados. A EMASESA efetuou inspeções *in loco* aos clientes selecionados pelos algoritmos em várias zonas da província de Sevilha, e segundo os autores a taxa de sucesso foi de aproximadamente 9%, considerada satisfatória pela natureza do problema, em que a fraude é fácil de esconder e raramente detectada.

Morote e Hernández-Hernández (2018) buscaram caracterizar os principais fatores que determinavam a ocorrência e a localização espacial das fraudes no sistema de distribuição de Alicante na Espanha. Segundo estes autores as maiores ocorrências de fraudes se concentram em regiões com menores rendas. Resultado semelhante foi obtido por Seago e McKenzie, (2007).

Al-Radaideh e Al-Zoubi (2018) desenvolveram um modelo indicador de fraudes para o SAA da *Yarmoukk Water Company* (YWC). Foram empregadas técnicas inteligentes de mineração de dados e dos algoritmos SVM e K-ésimo vizinho mais próximo (KNN) visando detectar essas atividades fraudulentas, reduzindo assim as perdas aparentes. A precisão do modelo gerado atingiu uma taxa de mais de 74% de precisão na identificação de fraudes o que é melhor do que os atuais procedimentos de predição manual feitos pelo YWC. Um estudo similar foi realizado para caracterizar os consumos fraudulentos de água usando dados históricos de consumo de 97.000 consumidores da empresa ESSBIO S.A., uma das principais concessionárias do Chile. Neste estudo foram empregados as técnicas de mineração de dados e os algoritmos Árvores de Decisão (CART), Naive Bayes (NB), Neuronal Net (NNet), SVM e KNN. O melhor desempenho foi obtido pela árvore de decisão seguida pelo SVM [Trancoso *et al.*, 2020].

Nofal *et al.* (2021) aplicaram vários métodos de detecção de anomalias a um conjunto de dados reais de hidrômetros mecânicos localizados na cidade de Amã, capital da Jordânia. O conjunto de dados empregados não é rotulado de forma que nenhuma discriminação é fornecida para qualquer medidor, independentemente de registrar um consumo normal de água ou não. Foram empregados diversos métodos de detecção de anomalias como a pontuação Z-score (ZS) [Fettermann *et al.*, 2015], fator de *outlier* local (LOF), agrupamento espacial baseado em densidade de aplicativos com ruído (DBSCAN), covariância mínima determinante (MCD), máquina de vetor de suporte de uma classe (OCSVM) e floresta de isolamento (iForest). Segundo estes autores o ZS, LOF, OCSVM e iForest sustentam a hipótese dos experimentos, onde 5, 10 e 16% do conjunto de dados são anômalos. Já os resultados do DBSCAN e MCD contrastam estas hipóteses.

3. MATERIAIS E MÉTODOS

3.1. ÁREA DE ESTUDO E DADOS

O município está localizado na região central da Região Metropolitana de Belo Horizonte. Segundo dados do último censo demográfico, realizado pelo Instituto Brasileiro de Geografia e Estatística (2021) em 2010, e obtido no portal Cidades@, a população do município era de 296.317 habitantes, com uma densidade demográfica de 1.905.07 hab./Km^2. O índice de Desenvolvimento Humano (IDH) era de 0.684, valor considerado médio, enquanto o IDH do Brasil era de 0.727 (alto) e o de Belo Horizonte era de 0.810 (muito alto). Ainda, segundos dados do instituto, o salário médio mensal era de 1.9 salários mínimos e apenas 9.2% da população total possuía ocupação. Na comparação com os outros municípios do estado, ocupava as posições 164 e 665 de 853, respectivamente. Considerando os domicílios com rendimentos mensais de até meio salário mínimo por pessoa, 34.5% da população se encontram nessas condições, o que o colocava na posição 555 dentre as cidades do estado.

Em relação às fraudes no SAA o percentual em janeiro de 2019 era de 45.36%. Em setembro de 2021 esse percentual saltou para 59.81%, evidenciando um acréscimo de 31.86%.

Os conjuntos de dados empregados neste estudo foram gerados pelo Sistema Comercial (SICOM) da Companhia de Saneamento de Minas Gerais (COPASA MG). Este é usado principalmente para emitir as faturas de água dos clientes, para gerir os cadastros dos consumidores e dos hidrômetros, além de gerir as ordens de serviço (O.S.). Os dados estão divididos em duas bases, onde a primeira contém as variáveis (atributos) de cadastro dos clientes, dos hidrômetros, dos volumes medidos e faturados (11/2019 a 05/2021) e a segunda contém os atributos das O.S. 152 de confirmação de infração (12/2020 a 03/2021).

Após a abertura da O.S. 152 (confirmação de infração) um funcionário vai a campo para checar se há ou não fraude. Após a inspeção a O.S. receberá um dos códigos de resultado de baixa, quais sejam: B01, B02, B03, B04, B05, F01 e F02 (infração confirmada, infração não confirmada, autodenuncia confirmada, autodenuncia não confirmada, aguardando regularização do usuário, furto/extravio confirmado e furto/extravio não confirmado, respectivamente). Ressaltamos que os códigos B01, B03 e F01 serão rotulados como fraude e recebem o rótulo 1 (fraude), os outros códigos recebem o rótulo 0 (não fraude).

As empresas de saneamento podem usar registros da localização dos pontos de serviços – zona de abastecimento (ZA), setor/rota (SR) e face (FAC) – da categoria tarifária (Residencial/social, comercial, industrial, pública e mista), das características dos equipamentos de micromedição (hidrômetros) e de dados coletados mensalmente

pelos leituristas utilizando o sistema de leitura e impressão de conta (SILEIM). Estes dados podem ser usados para criar perfis espaciais de possíveis fraudadores, visando otimizar as fiscalizações através da abertura da ordem de serviço (O.S) 152 (confirmação de infração), e consequentemente reduzindo o consumo não autorizado. Esse pode ser considerado um projeto especial de ciência de dados, onde a concessionária implanta um produto baseado em DM que pode ser potencialmente enriquecido por meio de diferentes novas explorações dos dados.

3.2. DATA SCIENCE TRAJECTORY (DST):

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

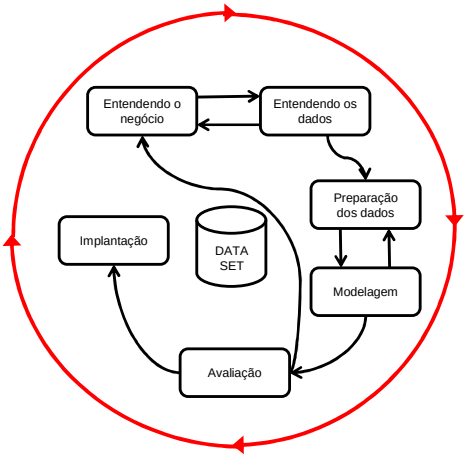


Figura 1 – Método de mineração de dados CRISP-DM [Adaptado Wirth e Hipp, 2000].

Qualquer projeto de mineração de dados (DM) começa com a definição da meta do projeto (objetivo) que está incluída na compreensão do negócio. Na fase de compreensão dos dados as hipóteses de informações ocultas relacionadas ao objetivo são formadas com base na experiência e em suposições qualificadas. Na fase de preparação os dados relevantes são coletados e pré-processados. Na fase de modelagem, um fluxo de trabalho de DM é construído para encontrar as configurações desejadas de parâmetros para os algoritmos selecionados e para executar a tarefa de mineração de dados nos dados pré-processados. Na fase de avaliação o modelo treinado é testado em relação a um conjunto de dados reais e os resultados da DM são avaliados de acordo com o objetivo de negócio. O conjunto de teste segue as etapas desenvolvidas nas fases de preparação e de modelagem. Após a avaliação do modelo estudado, tem-se a sexta, e última fase, que é pôr em produção (implantação).

A principal diferença entre a mineração de dados há vinte anos e a ciência de dados hoje é que a primeira é orientada a objetivos e se concentra no processo, enquanto a segunda é orientada a dados e é exploratória [Martínez-Plumed *et al.*, 2021]. Nesse contexto Martínez-Plumed *et al.* (2021) propuseram o novo modelo de trajetórias de ciência de dados denominado DST. Este modelo representa uma revisão importante do CRISP-DM original, visto que ele ainda representa uma das trajetórias mais comuns na ciência de dados, que vão dos dados ao conhecimento quando há um objetivo de negócio claro que se traduz em um objetivo de DM. Contudo, a DST permite a flexibilidade que a ciência de dados do século XXI exige, podendo incorporar as metodologias novas no desenvolvimento e implantação de projetos de ciência de dados [Martínez-Plumed *et al.*, 2021].

A ciência de dados é fundamentalmente exploratória e pode incluir algumas das seguintes atividades: exploração de objetivos de negócios que podem ser alcançados de forma orientada por dados; exploração de fonte de dados; exploração de valores que podem ser extraídos dos dados; exploração de resultados da ciência de dados aos objetivos do negócio; exploração narrativa que extraem histórias valiosas dos dados; e exploração de produto onde são encontradas novas maneiras de transformar o valor extraído dos dados em um serviço que forneça algo novo e valioso para usuários e clientes [Martínez-Plumed *et al.*, 2021].

Um projeto de ciência de dados bem-sucedido segue uma trajetória através de um espaço como o retratado em Figura 2. Diferentemente do CRISP-DM, a DST não possui setas dado que as atividades não devem ser realizadas em qualquer ordem pré-determinada [Martínez-Plumed *et al.*, 2021], sendo de responsabilidade do líder (es) de projeto decidir qual passo será dado a seguir. Mesmo que o mapa DST contenha todas as fases do CRISP-DM, estes não são necessariamente executados na ordem padrão. As atividades orientadas para o objetivo são intercaladas com atividades exploratórias.

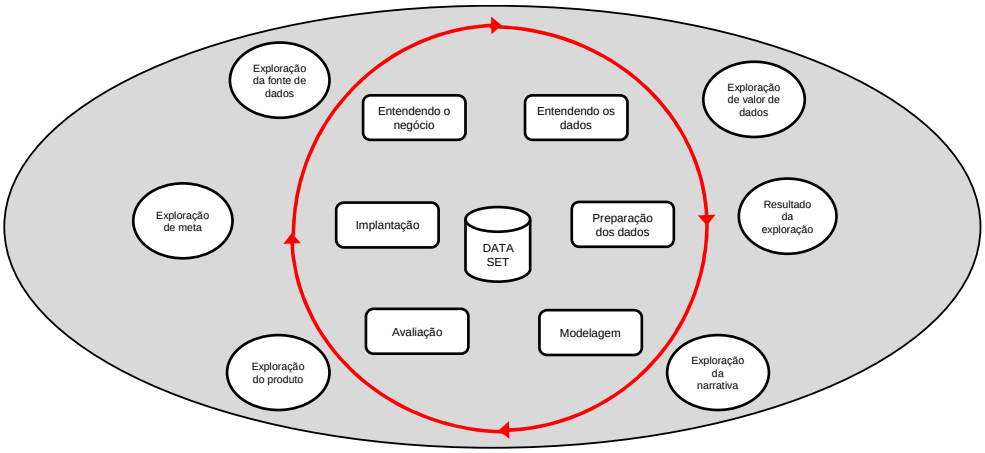


Figura 2 – O mapa de DST, contendo o círculo externo de atividades exploratórias, o círculo interno de atividades CRISP-DM (ou direcionadas a um objetivo) e, no centro, as atividades de gerenciamento de dados [Adaptado de Martínez-Plumed *et al.*, 2021].

A figura 3 representa a trajetória através do mapa DST onde a meta é estabelecida como uma primeira etapa de forma orientada a dados (exploração de meta) e dados relevantes são então explorados para extrair conhecimento valioso (exploração de valor de dados). As atividades clássicas do CRISP-DM são realizadas para limpar e transformar os dados (transformação de dados) que serão usados para treinar um determinado modelo de aprendizado de máquina (modelagem). Finalmente, o produto e/ou apresentação do usuário final mais apropriado é explorado (exploração do produto), a fim de transformar o valor extraído dos dados em um produto valioso para os usuários.

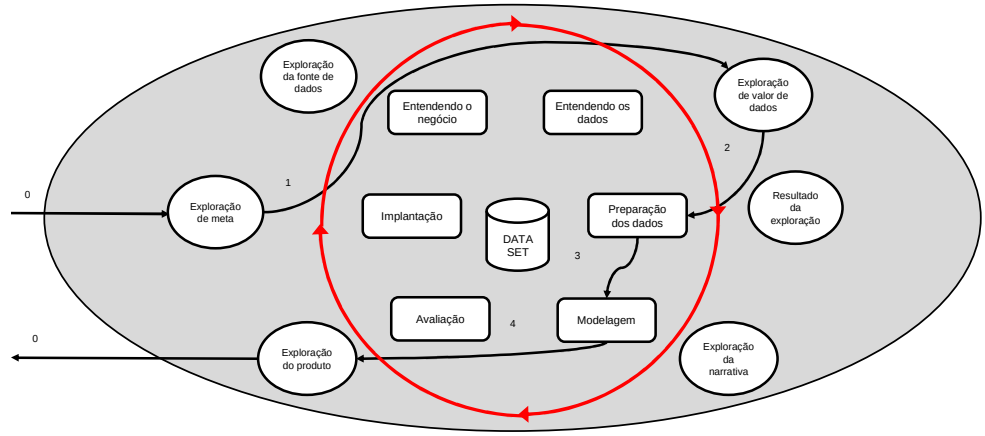


Figura 3 – Exemplo de trajetória através de um projeto de ciência de dados [Adaptado de Martínez-Plumed *et al.*, 2021].

As empresas de saneamento podem usar registros da localização dos pontos de serviços – zona de abastecimento (ZA), setor/rota (SR) e face (FAC) – da categoria tarifária (Residencial/social, comercial, industrial, pública e mista), das características dos equipamentos de micromedição (hidrômetros) e de dados coletados mensalmente pelos leituristas utilizando o sistema de leitura e impressão de conta (SILEIM). Estes dados podem ser usados para criar perfis espaciais de possíveis fraudadores, visando otimizar as fiscalizações através da abertura da ordem de serviço (O.S) 152 (confirmação de infração), e consequentemente reduzindo o consumo não autorizado. Esse pode ser considerado um projeto especial de ciência de dados, onde a concessionária implanta

um produto baseado em DM que pode ser potencialmente enriquecido por meio de diferentes novas explorações dos dados.

3.3. EXTREME GRADIENT BOOSTING (XGBOOST)

O XGBoost é a abreviação de eXtreme Gradient Boosting, algoritmo proposto por Chen e Guestrin, (2016). Esse algoritmo é uma implementação eficiente e escalonável da estrutura de aumento de gradiente proposto por Friedman, (2001) e Friedman *et al.*, (2000), e tem se tornando cada vez mais popular, não apenas para classificação, mas também para problemas de regressão devido ao seu alto desempenho [e.g. Chen e Guestrin, 2016; Lei *et al.*, 2019; Li *et al.*, 2020; Madhuri *et al.*, 2021]. A computação paralela, distribuída, out-of-core com reconhecimento de cache torna o algoritmo mais de dez vezes mais rápido do que os modelos populares usados no aprendizado de máquina e de aprendizado profunda. Ainda, segundo Chen e Guestrin, (2016), o algoritmo fornece resultados de última geração em muitos problemas de desafios de mineração de dados.

Deixe a saída de uma arvore ser

$$f(x) = w_q(x_i) \tag{1}$$

em que x é o vetor de entrada e w_q é o score correspondente da folha q. A saída de um conjunto de árvores k será

$$y_i = \sum_{k=1}^{k} f_k(x_i) \tag{2}$$

O algoritmo XGBoost tenta minimizar a função objetivo J na etapa t

$$J(t) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega\left(f_i\right) \tag{3}$$

em que o primeiro termo contém a função de perda do treino L (por exemplo, erro quadrático médio) entre a classe real y e a saída $\hat{y}$ para as n amostras e o segundo termo é o termo de regularização, que controla a complexidade do modelo e ajuda a evitar o *overfitting*.

No XGBoost, a complexidade é definida como:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

em que T é o número de folhas, γ é a pseudo-regularização do hiperparâmetro, dependendo de cada conjunto de dados e λ é a norma L_2 para os pesos das folhas.

Usando gradientes para a aproximação de segunda ordem da função de perda e encontrar os pesos ideais de w, o valor ideal da função objetivo é:

$$J(t) = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda} + \gamma T \tag{5}$$

em que $g_i = \partial_{\hat{y}^{t-1}} L(y, \hat{y}^{t-1})$ e $h_i = \partial^2{}_{\hat{y}^{t-1}} L(y, \hat{y}^{t-1})$ são as estatísticas de gradiente sobre a função de perda e I é o conjunto de folhas.

Os parâmetros de ajuste mais importantes do XGBoost são: (1) **nrounds**, que é o número máximo de iterações de boost. (2) **max_depth** é a profundidade máxima de uma árvore individual. (3) **min_child_weight** é a soma mínima do peso da instância necessária em um nó folha. (4) **gama** é a redução de perda mínima necessária para fazer uma partição adicional em um nó folha da árvore. (5) **subsample** é a proporção de subamostra das instâncias ou linhas de treinamento. (6) **learning_rate** é usado durante a atualização para evitar overfitting. O ajuste do XGBoost é complicado pois a alteração de qualquer um dos parâmetros pode afetar os valores ideais dos outros. Isto posto, a maioria dos estudos usa o valor padrão dos parâmetros para modelagem, e poucos estudos descreveram os detalhes do processo de ajuste dos parâmetros do XGBoost [Li *et al.*, 2020].

A abordagem de pesquisa em grade, usada para encontrar a melhor combinação de parâmetros, foi a mesma empregada por Li *et al*. (2020). A faixa de parâmetros para procurar o melhor conjunto de combinação é a seguinte: max_depth é de 2 a 10 com 2 etapas, min_child_weight é de 1 a 5 com 1 etapa, gamma é de 0 a 0,4 com 0,1 etapa, subamostra é de 0,6 a 1 com 0,1 passo e os valores da taxa de aprendizagem são 0,01, 0,05, 0,1, 0,2 e 0,3.

3.4. SHAPLEY ADDITIVE EXPLANATIONS (SHAP)

Segundo Li (2022), a inteligência artificial (AI) e o ML, anteriormente considerados, abordagens *black box*, estão se tornando mais interpretáveis, como resultado dos recentes avanços em *eXplainable AI* (XAI) [maiores detalhes ver Gunning e Aha, 2019], em particular, a interpretação local através de métodos como SHAP.

O SHAP é um dos métodos de atribuição de features aditivas de classe e é baseado na unificação da teoria dos jogos [Shapley, 1953], que visa distribuir de forma justa as contribuições dos jogadores quando eles alcançam coletivamente um determinado resultado e explicações locais [Lundberg e Lee, 2017]. Os valores de Shapley podem ser usados em ML para quantificar a contribuição de cada feature no modelo que fornece coletivamente a previsão [Strumbelj e Kononenko, 2014]. SHAP cria entradas simplificadas z' mapeando x para z' através de $x = h_x(z')$. Com base em z', o modelo original $f(x)$ pode ser aproximado com uma função linear de variáveis binárias:

$$f(x) = g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z_i' \tag{6}$$

em que $z' \in \{0,1\}^M$, M é o número de *features* de entrada, $\emptyset_0 \in \mathbb{R}$, o z_i' representa as *features* observadas ($z_i' = 1$) ou desconhecidas ($z_i' = 0$), e $\emptyset_i$'s são os valores atribuídos as *features*.

De acordo com Lundberg *et al.* (2019), o SHAP é o único método que possui todas as três propriedades desejáveis: precisão local, ausência e consistência e, portanto, é uma forte motivação para usar valores SHAP para atribuição das features do conjunto de árvores.

Para calcular valores SHAP:

$$f_x(S) = f\big(h_x(z')\big) = E[f(x)|x_s] \tag{7}$$

em que S é o conjunto de índices diferentes de zero em z' e $E[f(x)|x_s]$ é o valor esperado da função condicionada a um subconjunto S das *features* de entrada.

Os valores SHAP combinam essas expectativas condicionais com os valores Shapley clássicos da teoria dos jogos para atribuir valores ϕ_i a cada *feature*:

$$\emptyset_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \tag{8}$$

em que N é o conjunto de todos os *features* de entrada

O método Tree SHAP, usado neste trabalho, aproveita estruturas de árvore de decisão para calcular o valor $E[f(x)|x_s]$ de forma eficiente. Dado o número de árvores T e o número máximo de folhas em qualquer árvore L, a complexidade original do cálculo de $E[f(x)|x_s]$ é $O(TL2^M)$. Dada a profundidade máxima de qualquer árvore D, a complexidade de calcular $E[f(x)|x_s]$ com Tree SHAP é $O(TLD^2)$, o que diminui a complexidade computacional de um nível exponencial de alta ordem para um nível quadrático [Lundberg e Lee, 2019].

O Tree SHAP é um modelo baseado em árvore treinado com os dados de entrada X usados para treinar o modelo ($N \times M$ matriz de N instâncias e M features). Estes geram uma matriz $N \times M$ dos valores SHAP. Cada valor indica o quanto a feature contribui para a previsão da instância correspondente.

4. RESULTADOS E DISCUSSÕES

A base de dados compreende o período de 11/2019 a 05/2021. Por se tratar de um banco de dados desbalanceado, foi empregado técnicas de balanceamento e este conjunto de dados foi dividido em em 70% para treino e 30% para teste.

Para avaliar a precisão dos resultados obtidos através dos modelos classificadores empregados, foram utilizados três medidas normalmente utilizadas no meio científico, quais são: Acurácia, coeficiente de Kappa[1] e área sobre a curva (AUC). Estas medidas estão relacionadas com os resultados de classificação e diagnóstico verdadeiro.

O teste é considerado *"yes"* (fraude) ou *"no"* (não fraude). O teste está correto quando ele é *"yes"* na presença de fraude (VP) ou *"no"* na presença de não fraude (VN). Além disso, o teste está errado quando ele é *"yes"* na ausência de fraude (FP), erro tipo I, ou *"no"* quando a fraude está presente (FN), erro tipo II. Esses quarto valores compõem a chamada matriz de confusão.

$$MC = \begin{pmatrix} VN & FN \\ FP & VP \end{pmatrix}$$

A acurácia do modelo foi de 0,997, indicando a sua capacidade em identificar corretamente se há ou não fraude, o coeficiente Kappa foi de 0,772, demonstrando grau moderado de nível de concordância ou reprodutividade e a AUC foi de 0,994. Segundo Cohen (1960), quando a AUC estiver entre 0,70 e 0,80, o modelo tem poder discriminatório aceitável.

A Curva ROC, representada pela Figura 4, consiste em um gráfico de duas dimensões onde o eixo *"y"* refere-se ao Recall/Precisão, representando a taxa de verdadeiros positivos (TVP), e o eixo *"x"* a especificidade, representando a taxa de falsos positivos (TFP).

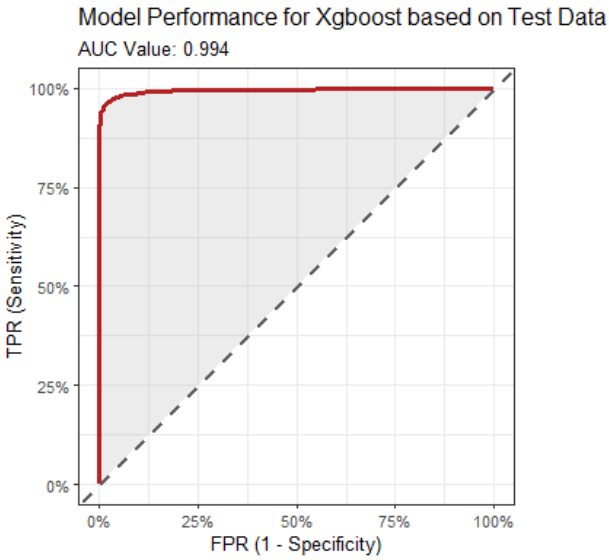


Figura 4 – Curva ROC

Com o objetivo de inferir a importância das *features* que possuem um maior peso na detecção de fraude, empregou-se o método SHAP. Este método atribuirá pesos diferentes as *features*, conforme sua percepção de padrão de acontecimentos da *features* preditiva na classificação. Os resultados estão apresentados nas Figuras 5 e 6.

[1] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 1, 37-46

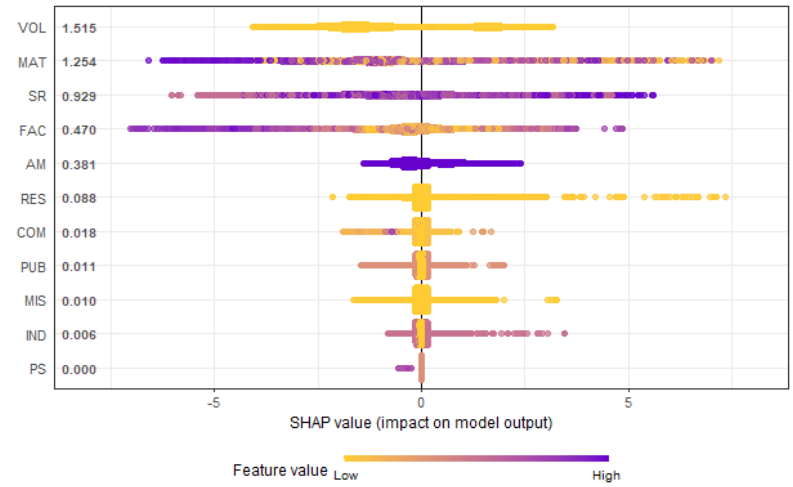


Figura 5 – Importância do SHAP para o modelo XGBoost.

A Figura 5 mostra a distribuição dos valores SHAP de todas as instâncias para cada *features*, mostrando o impacto individual e orientado (positivo ou negativo) das instâncias na classificação. A *feature* mais influente é o volume, seguido pela matrícula, pelo setor rota e pela face.

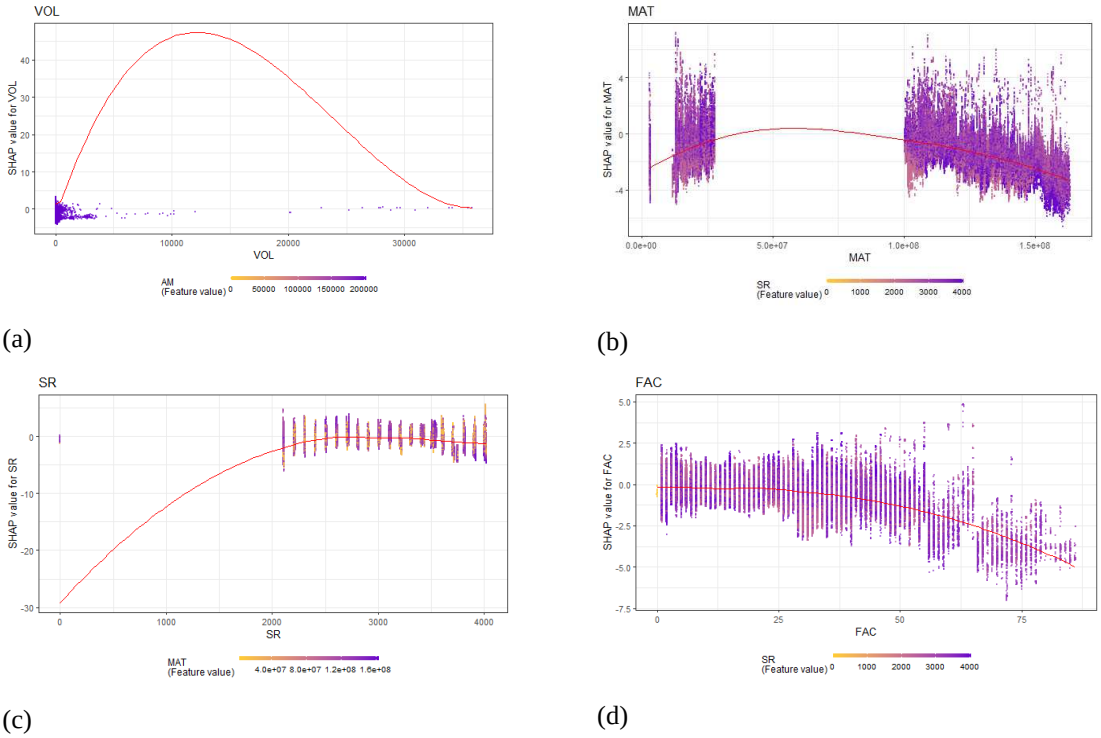


Figura 6 – Gráficos de dependência SHAP das quarto principais *features*

A Figura 6 apresenta uma representação mais detalhada da dependência das *features* no modelo XGBoost considerando cada ponto de dados no conjunto de dados individualmente. Os valores das *features* e os valores SHAP correspondentes são plotados nos eixo x e y, respectivamente. Deve-se notar que os valores SHAP não representam relações causais, mas descrevem o comportamento do modelo. A Figura 6(d) demonstra uma relação inversa entre fraudes e face. Valores mais altos de fraude podem ser obtidos quando a face está entre 0 e 25 e o setor rota acima de 3000.

A DST permite a flexibilidade que a ciência de dados do século XXI exige [Martínez-Plumed *et al.*, 2021], podendo incorporar metodologias no desenvolvimento e implantação dos projetos. Visando caracterizar

espacialmente as fraudes, empregando a logística comercial de leitura de hidômetros, incorporou-se a técnica InfoVis, empregada na construção do *Treemap* [Shneiderman, 1992].

Visando validar esse conceito foi definido dois setores/rotas e suas respectivas faces para a realização de uma inspeção *in loco*. A Figura 7 apresenta, para cada setor/rota, quais as faces têm maior probabilidade de ocorrerem fraudes. Esta técnica, possibilita a visualização da estrutura hierárquica de um diagrama de árvore, representado pelos valores quantitativos dos dados através do tamanho da área do quadrante. No SR 2714, as faces com maior probabilidade de ocorrerem fraudes são: 9, seguida pela 2, pela 11, e assim sucessivamente. O mesmo racional é aplicado para o SR 2715. O percentual de sucesso da inspeção *in loco* foi de 11,7%, resultado considerado satisfatório pela natureza do problema, em que a fraude é fácil de esconder e raramente detectada.



Figura 7 – Setor Rota/Face com maior probabilidade de ocorrer fraudes.

5. CONCLUSÕES

O estudo realizado propõe uma nova sistemática para detecção dos locais das fraudes nos sistemas de abastecimento de água (SAA) aplicando a logística comercial de leitura dos hidrômetros, juntamente com o DST, o algoritmo XGBoost e o método SHAP. A escolha do XGBoost decorreu por ser uma implementação eficiente e escalonável da estrutura de aumento de gradiente que vem sendo empregado, não apenas para classificação, mas também para problemas de regressão devido ao seu alto desempenho. Já o emprego do SHAP visa aprimorar a explicabilidade do algoritmo de ML, avaliando as influências das *features*, as dependências e as interações com a fraude. A definição de quais *features* são mais importantes dão o direcionamento a novos caminhos que essa pesquisa pode seguir.

A idéia subjacente é explicar a contribuição de todas as *features* na previsão identificando a dependência e as interações entre *features*, garantindo que os detalhes implícitos sejam descobertos e examinados minuciosamente, permitindo selecionar um modelo genuinamente confiável e robusto para interpretação. O uso de um método de explicação preciso, robusto e granular, como o SHAP, pode gerar *insights* extras tanto na comparação entre modelos quanto na sua interpretação, tendo potencial de ser aplicado para quaisquer fins.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. Al-Radaideh, Q.A., Al-Zoubi, M.M. *A data mining based model detection for fraudulent behavior in water consumption. In: 9th International Conference on Information and Communication Systems (ICICS)*, 2018.

2. Alegre, H., Hirner, W., Baptista, J.M., Cabrera Jr., E., Cubillo, F., Duarte, P., Merkel, W., Parena, R. *Performance indicators for water supply services.* 3nd edition; IWA Publishing: Londres, UK, 2016.

3. Bharti, N., Khandekar, N., Sengupta, P., Bhadwal, S., Kochhar, I. *Dynamics of urban water supply management of two Himalayan towns in India. Water Policy*, 22, 65-89, 2000.

4. Biswas, A.K., Tortajada, C. *Water Supply of Phnom Penh: An Example of Good Governance. International. Journal of Water Resources Development*, 26, 2, 157-172, 2010.

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

6. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y. *xgboost: Extreme Gradient Boosting.* 2021. Disponível em: <https://cran.r-project.org/web/packages/xgboost/index.html>. Acesso em: 04 Novembro 2021.

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

8. Cohen, J. A *coefficient of agreement for nominal scales. Educational and Psychological Measurement*, 20(1), 37-46, 1960.

9. Fanner, P., Thornton, J., Liemberger, R., Sturm, R. *Evaluating water loss and planning loss reduction strategies; American Water Works Research Foundation*, 2007.

10. Farley, M., Liemberger, R. *Developing a non-revenue water reduction strategy: planning and implementing the strategy. Water Science and Technology. Water Supply*, 5(1), 41-50, 2005.

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

12. Faria, L.T., Melo, J.D., Padilha-Feltrin, A. *Análise Espacial de Pontos para Mapeamento de Perdas Comerciais. In: V Simpósio Brasileiro de Sistemas Elétricos – V SBSE*, Foz do Iguaçu-PR, 22 a 25 de abril de 2014, 1-6.

13. Fettermann, D.C., Guerra, K.C., Mano, A.P., Marodin, G.A. *Uma sistemática para detecção de fraudes em empresas de abastecimento de água. Interciência*, 40(2), 114-121, 2015.

14. Friedman, J.H., Hastie, T., Tibshirani, R. *Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics*, 28(2), 337-407, 2000.

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

16. Garcia, D., Gonzalez, D., Quevedo, J., Puig, V., Saludes, J. *Water demand estimation and outlier detection from smart meter data using classification and Big Data methods. In: Proceedings of 2nd IWA New Developments in IT and Water Conference:* Rotterdam, Holland, 8-10 February 2015.

17. Gonzáles-Gómes, F., Martinez-Espineira, R., Garcia-Valiñas, M.A., García-Rubio, M.A. *Explanatory factors of urban water leakage rates in southern Spain. Urban Policy*, 22, 22-30, 2012.

18. Gunning, D.; Aha, D. *DARPA's explainable artificial intelligence (XAI) program. AI Magazine*, 38(3), 50-57, 2019.

19. Henriques, J., Caldeira, F., Cruz, T., Simões, P. *Combining K-Means and XGBoost models detection using log datasets. Eletronics*, 9(1164), 1-16, 2020.

20. Humaid, E.H., Barhoum, T. *Water consumption financial fraud detection: a model based on rule induction. In: Palestinian International Conference on Information and Communication Technology Gaz*, Palestin, 15-16 April, 2013, 115-120, 2013.

21. IBGE – Instituto Brasileiro de Geografia e Estatística. Cidades@. 2021. Disponível em: <https://cidades.ibge.gov.br/brasil/mg/ribeirao-das-neves/panorama>. Acesso em: 20 Dezembro 2021.

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

23. Jung, D., Kang, D., Liu, J., Lansey, K. *Improving the rapidity of responses to pipe burst in water distribution systems: A comparison of statistical process control methods. Journal of Hydroinformatics*, 17(2), 307- 328, 2015.

24. Kainz, O., Karpiel, E., Petija, R., Michalko, M., Jakab, F. *Non-standard situation detection in smart water metering. Open Computational Science*, 11, 12-21, 2021.

25. Kingdom, B., Liemberger, R., Marin, P. *The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries - How the Private Sector Can Help: A Look at Performance-Based Service Contracting.* Water Supply and Sanitation Board Discussion Paper Series, Washington, 2006, 8, 1-40. Disponível em: <https://ppiaf.org/documents/2076/download>. Acesso em: 03 Novembro 2021.

26. Lambert, A., Hirner, W. *Losses from Water Supply Systems: Standart Termonology and Recommended Performance Measures.* IWA- International Water Association. U.K., 4-7, 2000.

27. Lai, C.H, Chan, N.W., Roy, R. *Understanding public perception of and participation in non-revenue water management in Malaysia to support urban water policy. Water*, 9(26), 1-16, 2016.

28. Lei, Y., Jiang, W., Jiang, A., Zhu, Y., Niu, H., Zhang, S. *Fault diagnosis method for hydraulic directional valves integrating PCA and XGBoost. Processes*, 7(589), 1-12, 2019.

29. Li, Y., Li, M., Li, C., Liu, Z. *Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Scientific Reports*, 10, 9952, 2020.

30. Li, Z. *Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Computers, Environment and Urban Systems*, 96, 101845, 2022.

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

32. Liu, Y., Just, A. *SHAPforxgboost: SHAP Plots for 'XGBoost'*. R package version 0.1.0., 2020. https://github.com/liuyanguu/SHAPforxgboost/

33. Lundberg, S.M., Lee, S.I. *A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems*, 30, 4768-4777, 2017.

34. Lundberg, S.M., Erion, G.G., Lee, S.I. *Consistent individualized feature attribution for tree ensembles*. arXiv, 2019, 1802.03888v3.

35. Ma, J., Ding, Y., Cheng, J.C.P., Tan, Y., Gan, V.J. L., Zhang, J. *Analyzing the leading causes of traffic fatalities using XGBoost and Grid-Based analysis: A city management perspective. IEEE Access*, 7, 148059-148072, 2020.

36. Maddocks, A., Young, R.S., Reig, P. *Ranking the World's Most Water-Stressed Countries in 2040*. World Resources Institute 2015. Disponível em: <http://www.wri.org/blog/2015/08/ranking-world%E2%80%99s-most-water-stressed-ountries-2040>. Acesso em: 24 Junho 2018.

37. Madhuri, R., Sistla, S., Raju, K.S. *Application of machine learning algorithms for flood susceptibility assessment and risk management. Journal of Water and Climate Change*, 12(6), 2608-2623, 2021.

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

39. Mashford, J., De Silva, D., Burn, S., Marney, D. *Leak detection in simulated water pipe networks using svm. Applied Artificial Intelligence* 2012, 26, 5, 429-444.

40. Mcdonald, R.I., Weber, K., Padowski, J., Flörke, M., Schneider, C., Green, P.A., Gleeson, T., Eckman, S., Lehner, B., Balk, D., Boucher, T., Grill, G., Montgomery, M. *Water on an urban planet: Urbanization and the reach of urban water infrastructure. Global Environmental Change*, 27, 96-105, 2014.

41. Meseguer, J., Mirats-Tur, J.M., Cembrano, G., Puig, V., Quevedo, J., Pérez, R., Sanz, G., Ibarra, D. *A decision support system for on-line leakage localization. Environmental Modelling Software*, 60, 331-345, 2014.

42. Molinos-Senante, M., Mocholí-Arce, M., Sala-Garrido, R. *Estimating the environmental and resource costs of leakage in water in a distribution systems: a shadow price approach. The Science of the Total Environment*, 568(15), 180-188, 2016.

43. Monedero, I., Biscarri, F., Guerrero, J.I., Roldán, M., León, C. *An approach to detection of tampering in water meters. Procedia Computer Science*, 60, 413-421, 2015.

44. Monedero, I., Biscarri, F., Guerrero, J.I., Peña, M., Roldán, M., León, C. *Detection of water meter under-registration using statistical algorithms. Journal of Water Resources Planning and Management* 2016, 142, 1.

45. Morote, A.F., Hernández-Hernández, M. Unauthorised domestic water consumption in the city of Alicante (Spain): A consideration of its causes and urban distribution (2005-2017). Water, 10(7), 851, 2018.

46. Mounce, S.R., Mounce, R.B., Boxall, J.B. *Novelty detection for time series data analysis in water distribution systems using support vector machines. Journal of Hydroinformatics*, 13(4), 672-686, 2011.

47. Mutikanga, H.E., Sharma, S.K., Vairavamoorthy, K. *Assessment of apparent losses in urban water systems. Water and Environment Journal*, 25(3), 327-335, 2011.

48. Nofal, S., Alfarrarjeh, A., Jabal, A.A. *A use of anomaly detection for identifying unusual water consumption in Jordan. Water Supply*, ws2021210, 2021.

49. Seago, C.L., Mckenzie, R.S. An *Assessment of Non-Revenue Water in South Africa. Pretória: Water Research Comission* - WRC Report No TT 300/0, 2007. 112p. Disponível em: <http://ws.dwa.gov.za/mwg-internal/de5fs23hu73ds/progress?id=Sb-n7bENvK6tnFIdstRlCdpwHQntaORxaSSxEN0y4Z8>. Acesso em: 03 Novembro 2021.

50. Shapley, L. *A Value for n-Person Games. In: Kuhn, H.; Tucker, A., Eds., Contributions to the Theory of Games II*, Princeton University Press, Princeton, 307-317, 1953.

51. Shneiderman, B. *Tree visualization with Tree-Maps: 2-d Space-filling approach. ACM Transactions on Graphics*, 1992, 11, 1, 92-99.

52. Srirangarajan, S., Allen, M., Preis, A., Iqbal, M., Hock Beng Lim, H.B., Whittle, A.J. *Wavelet-based Burst Event Detection and Localization in Water Distribution Systems. Journal of Signal Processing Systems*, 7(1), 1-16, 2013.

53. Strumbelj, E., Kononenko, I. *Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems*, 41(3), 647-665, 2014.

54. Trancoso, E.F.H., Figueroa, F., Gisselot, P. *Predicción de fraudes en el consumo de agua potable mediante el uso de minería de datos. Universid, Ciencia y Tecnología*, 24(104), 58-66, 2020.

55. Vörösmarty, C.J., Green, P.A., Salisbury, J., Lammers, R.B. *Global Water Resources: Vulnerability from Climate Change and Population Growth. Science*, 289(5477), 284-288, 2000.

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