

## 242 - WATER FRAUD ANALYTICS – UM MODELO DE MACHINE LEARNING PARA DETECÇÃO DE FRAUDES EM CONSUMO DE ÁGUA

### **Márcio Nunes de Souza<sup>(1)</sup>**

Graduado em Informática (UCSal). Mestre em Engenharia de Sistemas e Produtos (IFBA). Analista de Tecnologia da Informação (Embasa).

### **Renato Lima Novais<sup>(2)</sup>**

Graduado em Ciência da Computação (UFBA). Mestre em Informática (Puc-Rio). Doutor em Ciência da Computação (UFBA). Docente de Graduação e Pós-Graduação (IFBA).

### **Rodrigo Tripodi Calumby<sup>(3)</sup>**

Bacharel, Mestre e Doutor em Ciência da Computação. Docente de Graduação e Pós-Graduação em Ciência da Computação (PGCC/UEFS).

**Endereço<sup>(1)</sup>:** 4ª Avenida, 420, Centro Administrativo da Bahia – CAB - 41745-002 - Salvador - Bahia - Brasil -  
Tel: +55 (71) 99191-3313 - e-mail: mnunes.ti@gmail.com.

### **RESUMO**

As perdas de água nos sistemas de distribuição ocorrem frequentemente no setor de saneamento. Dentre os tipos de perdas, as fraudes correspondem ao volume de água furtado pelos usuários e, tradicionalmente, são combatidas por meio da inspeção da rede de abastecimento, que é uma atividade complexa e custosa, resultando em uma baixa taxa de mitigação e perdas financeiras para as empresas de água. Técnicas de aprendizado de máquina têm sido aplicadas para identificar fraudes em vários setores com resultados promissores. No entanto, no saneamento, estudos anteriores realizados apresentaram várias limitações em termos de volume de dados e rigor científico. Este trabalho apresenta um estudo de identificação de fraudes no consumo de água usando um conjunto de algoritmos de aprendizado de máquina supervisionado. Consideramos um banco de dados em larga escala e procedimentos experimentais focando na construção e avaliação de modelos preditivos e seu poder de generalização. Entre os algoritmos de aprendizado de máquina avaliados, o melhor modelo apresentou uma precisão geral de 79,62% e foi capaz de identificar corretamente mais de 76% das fraudes.

**PALAVRAS-CHAVE:** fraudes, saneamento, consumo de água, perdas de água, aprendizado de máquina.

## INTRODUÇÃO

As perdas de água na distribuição são uma preocupação mundial de longo prazo Hudson (1964); Mutikanga et al. (2011). No Brasil, o setor de saneamento sempre enfrentou um cenário desafiador devido às altas taxas de perdas de água tratada IFC (2013). De acordo com o Sistema Nacional de Informações sobre Saneamento (SNIS), a média nacional das perdas de água nos sistemas de distribuição (IN049) é de 38,5% Brasil (2019). Essas perdas equivalem a R\$ 10,5 bilhões em prejuízos financeiros, segundo estudos do Instituto Trata Brasil (Brasil, 2020). Enquanto isso, em países mais desenvolvidos, os níveis de perdas são abaixo de 20% Trata Brasil (2020). Isso motivou a recente aprovação do novo marco legal do saneamento básico, que inclui metas para a redução dessas perdas pelas concessionárias responsáveis pelo abastecimento de água.

Em termos práticos, grandes perdas de água levam à redução da receita, resultando em uma diminuição na capacidade das empresas de fazer investimentos para expansão e melhorias nos serviços. As perdas de água também podem causar danos ao meio ambiente devido à maior exploração das fontes de água IFC (2013); AL-Washali et al. (2020); Mutikanga et al. (2013).

Em geral, as perdas podem ser classificadas como perdas reais, causadas por vazamentos, ou perdas aparentes, causadas por falhas no cadastro de clientes, erros de medição ou fraudes por parte dos usuários Kusterko et al. (2018). Mais especificamente, fraude ou consumo não autorizado é o volume de água furtados pelos usuários, por exemplo, usando ligações clandestinas, ligações diretas (*by-pass*) ou violações dos medidores de consumo AESBE (2015). As empresas de abastecimento de água enfrentam dificuldades em identificar onde ocorrem fraudes ou vazamentos. Um dos sinais indicativos de possível fraude é uma variação abrupta no volume de água consumida por um cliente. No entanto, mesmo nesses casos, a identificação de fraudes continua sendo um desafio, devido à variabilidade natural do padrão de consumo, como aquela resultante de mudanças no número de residentes, fatores esporádicos ou fatores sazonais, entre outros De Castro Fettermann et al. (2015).

O principal método utilizado para identificar fraudes é a inspeção física de toda a extensão das tubulações, usando equipamentos acústicos para buscar um fluxo de água indevido. Considerando a grande extensão da rede de abastecimento, esta é uma tarefa comumente realizada por amostragem dos equipamentos a serem inspecionados. Mesmo quando há uma tentativa de selecionar ativamente áreas para inspeção, isso é frequentemente realizado sem base em dados Gumier et al. (2007).

Muitos métodos de identificação de fraudes aplicados em outras áreas de negócios também podem ser aplicados no setor de saneamento. A exemplo de crimes como lavagem de dinheiro, fraudes com cartão de crédito, fraudes em comércio eletrônico, fraudes em telecomunicações e *hacking* de computadores usam modelos estatísticos e de aprendizado de máquina Bolton & Hand (2002). No contexto do aprendizado de máquina, a classificação e agrupamento de clientes de acordo com seus perfis, previsão de consumo, identificação de anomalias e construção de redes de relacionamento são aplicações usadas em vários tipos de fraudes Baesens et al. (2015). Embora usada em muitos negócios, a descoberta de fraudes com algoritmos de aprendizado de máquina ainda é rara no setor de saneamento. Poucos estudos aplicaram aprendizado de máquina a esse problema ou alcançaram eficácia prática AL-Washali et al. (2016).

Passini & Toledo (2002) usaram modelos neurais e árvore de decisão para identificar fraudes no consumo de água em Campinas-São Paulo. De Castro Fettermann et al. (2015) usaram o algoritmo Z-Score para identificar fraudes em Jequié-Bahia. Humaid & Barhoum (2013) criaram um modelo de indução de regras para fraudes no consumo de água em Gaza-Palestina. Monedero et al. (2015) desenvolveram três modelos baseados em correlação de *Pearson* e conjuntos de regras em Sevilha-Espanha. Al-Radaideh & Al-Zoubi (2018) usaram os algoritmos SVM e KNN em Irbid-Jordânia. Uddin et al. (2019); Sriramulu et al. (2020); Gopal & Balaji (2020); Sreedevi & Swathi (2021) propuseram usar os algoritmos SVM e KNN para perdas não técnicas. Espinosa et al. (2020) usaram mineração de dados para prever fraudes no consumo de água. Sreekanth & Thinakaran (2021) aplicaram os algoritmos CNN e RNN para prever fraudes em redes metropolitanas. Em sua maioria, os trabalhos anteriores conseguiram demonstrar a aplicabilidade de algoritmos e técnicas estatísticas para mitigar as fraudes no consumo de água, apesar das limitações metodológicas ou resultados insatisfatórios.

Considerando os impactos causados pelas fraudes nos sistemas de abastecimento de água, a baixa eficácia dos métodos existentes e o rigor científico dos estudos anteriores, estudos futuros devem focar em soluções robustas para combater fraudes no sistema de abastecimento de água. Nesse contexto, este trabalho empreende o desenvolvimento e validação experimental rigorosa de modelos de aprendizado de máquina para a descoberta de fraudes no consumo de água,

visando alcançar o poder de generalização em um cenário de grande escala e viabilizar sua aplicação prática. Entre os algoritmos de aprendizado de máquina avaliados, o melhor modelo apresentou uma precisão geral de 79,62% e foi capaz de identificar corretamente mais de 76% das fraudes.

## OBJETIVOS

Para satisfazer à necessidade de inovar nos métodos de mitigação e prevenção de fraudes no consumo de água, o objetivo geral deste trabalho visa o desenvolvimento de um modelo analítico que usa *machine learning* para auxiliar na identificação de fraudes no consumo de água e, conseqüentemente, proporcionar a redução das perdas aparentes nos sistemas de abastecimento de água.

## METODOLOGIA UTILIZADA

O método proposto, desenvolvido com o objetivo de construir e avaliar modelos de aprendizado de máquina para identificar fraudes no consumo de água entre clientes de uma empresa pública de saneamento, representado na Figura 1, foi baseado no processo de Descoberta de Conhecimento em Bases de Dados (KDD) para projetos de mineração de dados de Fayyad et al. (1996). Todas as atividades foram realizadas usando a plataforma de análise de dados Knime KNIME (2020).

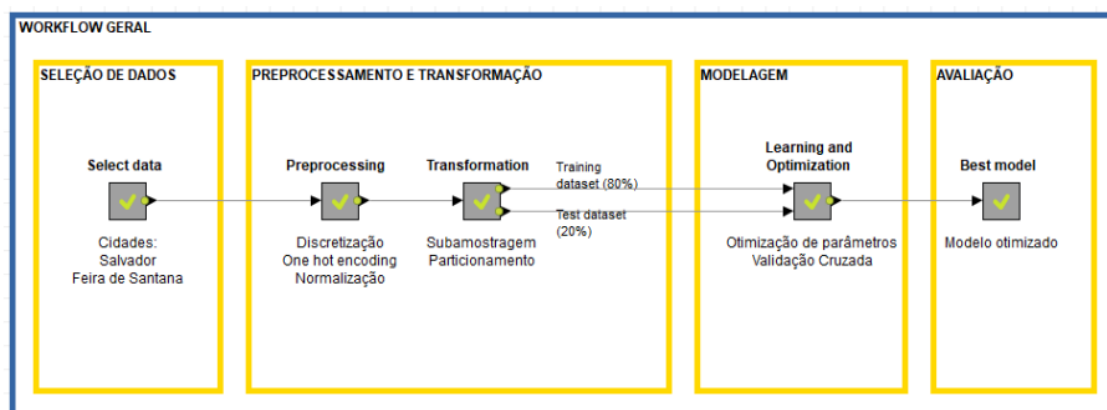


Figura 1: Workflow geral do processo para treinamento do modelo.

A partir do conjunto de dados original, selecionamos os dados a serem utilizados no processo experimental na etapa “Seleção de Dados”. Na etapa denominada “Pré-processamento e Transformação”, aplicamos técnicas de preparação, amostragem e particionamento de dados. Em seguida, na etapa “Aprendizado de Máquina”, realizamos as atividades de treinamento e otimização de algoritmos de aprendizado supervisionado. Finalmente, na etapa “Avaliação”, avaliamos os resultados dos algoritmos para apoiar a escolha para a implementação prática. As seções seguintes descrevem cada uma dessas fases em detalhes:

### Seleção de Dados

O conjunto de dados utilizado para treinar os modelos de aprendizado de máquina foram obtidos na Empresa Baiana de Águas e Saneamento S/A (Embasa) EMBASA (2020). Os dados incluem informações de cadastro de clientes - 4.390.219 matrículas – e os dados de consumo de água do ano de 2018. Em seguida, foram selecionadas as matrículas das cidades de Salvador e Feira de Santana, Bahia, Brasil. Além de serem as duas maiores cidades da Bahia, elas também são as que apresentam a maior incidência de fraudes identificadas, representando 53% do número total de fraudes identificadas em 2018 em todo o estado.

### Pré-processamento e Transformação

Foram aplicadas técnicas de pré-processamento para tornar os dados adequados para processamento com algoritmos de classificação supervisionada. Para as variáveis categóricas, aplicou-se a técnica “one hot encoding”, enquanto para as variáveis numéricas, a técnica de normalização Z-Score. A variável 'ConsumoMédio' foi discretizada conforme as nove categorias de consumo da Embasa.

No conjunto de dados, havia 30.003 matrículas com a identificação de fraudes anteriores, representando

aproximadamente 3% do total. Normalmente, conjunto de dados desbalanceados dificultam a aplicação de algoritmos de classificação Haixiang et al. (2017). Para equilibrar as classes, usou-se a técnica de subamostragem, eliminando exemplos da classe majoritária Thabtah et al. (2020). Assim, foram selecionados aleatoriamente 30.003 matrículas sem fraude e as 30.003 matrículas de fraudes existentes, totalizando 60.006 matrículas para o experimento.

Por fim, os dados foram particionados antes da etapa de modelagem. Destinou-se 80% do conjunto de dados para a otimização e treinamento dos classificadores, enquanto os 20% restantes foram reservados para o teste final de eficácia. Em todas as partições de dados realizadas, utilizou-se o método de amostragem estratificada com base na variável alvo. A **Erro! Fonte de referência não encontrada.** apresenta as tarefas realizadas nas etapas de pré-processamento e transformação dos dados na ferramenta Knime.

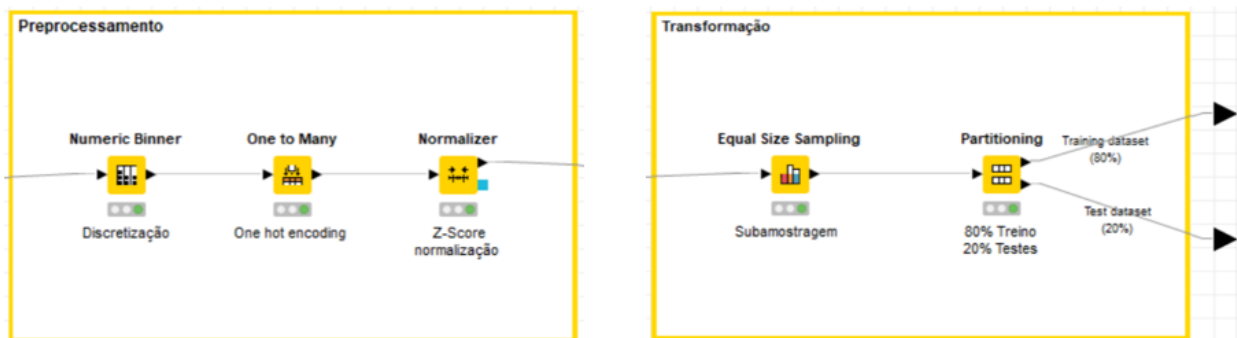


Figura 2: Pré-processamento e Transformação dos dados.

## Modelagem

No processo experimental de construção e validação dos preditores, utilizou-se algoritmos tradicionais de aprendizado supervisionado que se mostraram eficazes em diversas outras atividades Baesens et al. (2015). Entre eles: *Naive Bayes*, Regressão Logística, Árvores de Decisão, Máquina de Vetores de Suporte (SVM), *Random Florest* e *Gradient Boosting*. A Figura 3 apresenta a etapa de modelagem usando os algoritmos de classificação na ferramenta Knime.

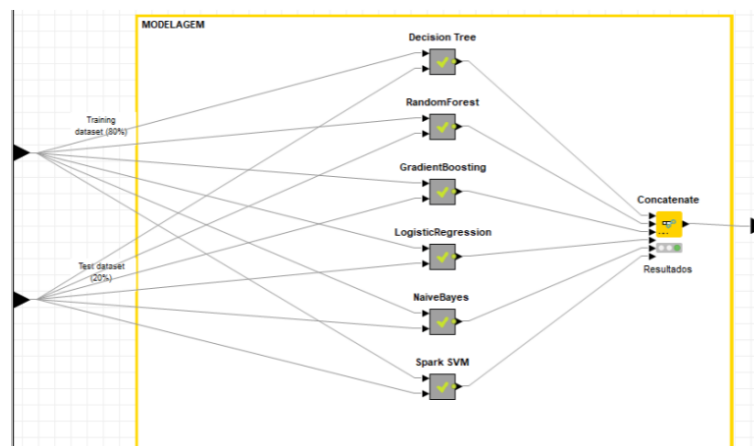


Figura 3: Etapa de modelagem com os algoritmos de classificação.

Para apoiar a escolha da melhor configuração para cada algoritmo, aplicamos a técnica de otimização de hiperparâmetros. Durante o processo de otimização, aplicou-se também a validação cruzada em dez partes. Assim, o conjunto de treinamento foi subdividido em 10 subconjuntos, sendo que em cada rodada, nove partes usadas para construir os modelos e uma parte usada para validação. Em cada iteração, um dos subconjuntos foi usado para teste e os demais para construção do modelo.

### Avaliação

Para cada iteração no conjunto de hiperparâmetros, foram gerados 10 resultados para realizar a validação cruzada. Assim, a média e o desvio padrão das medidas de avaliação tradicionais na área (*Recall*, *Precision*, *F-measure*, *Accuracy* e *Error*) foram calculados para obter os resultados. Para cada algoritmo, selecionamos a configuração de hiperparâmetros que apresentou os melhores resultados com base na média de acurácia. A Figura 4 apresenta o processo de otimização dos parâmetros e validação cruzada na ferramenta Knime.

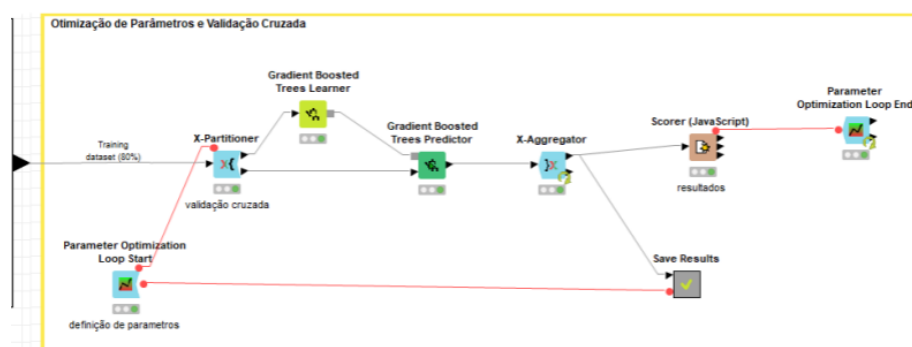


Figura 4: Processo de otimização e validação do modelo.

Escolheu-se as melhores configurações de hiperparâmetros para construir os modelos finais com todo o subconjunto de dados de treinamento. O subconjunto de dados de teste foi utilizado para a avaliação final dos modelos. Finalmente, o modelo que apresentou os melhores resultados foi selecionado, comparando os resultados de cada um dos modelos construídos.

### RESULTADOS E DISCUSSÃO

O processo de otimização paramétrica dos algoritmos de aprendizado foi utilizado para construir e avaliar os modelos, selecionando o mais eficaz para identificar fraudes no consumo de água. Calculou-se as principais medidas de avaliação dos modelos para apoiar a escolha das melhores configurações dos algoritmos. A medida de acurácia geral foi utilizada para escolher as melhores configurações e também para definir o melhor modelo. A **Tabela 1** apresenta os resultados obtidos por cada um dos modelos para o conjunto final de testes, em ordem de precisão. Pode-se observar que, entre os seis algoritmos avaliados, o *Random Forest* e o *Gradient Boosting* obtiveram resultados semelhantes, com uma ligeira superioridade para o *Gradient Boosting*.

Tabela 1: Resultados finais dos modelos.

Algoritmo	Verdadeiro Positivo	Falso Positivo	Verdadeiro Negativo	Falso Negativo	Acurácia
Gradient Boosting	76,34%	17,10%	82,90%	23,66%	79,62%
Random Florest	75,15%	17,41%	82,59%	24,85%	78,87%
Regressão Logística	69,91%	16,81%	83,19%	30,09%	76,55%
Linear SVM	71,04%	20,00%	80,00%	28,96%	75,52%
Naive Bayes	73,82%	25,35%	74,65%	26,18%	74,24%
Árvore de Decisão	70,97%	23,08%	76,92%	29,03%	73,95%

A Figura 5 apresenta a matriz de confusão e as estatísticas dos resultados obtidos pelo modelo treinado com o algoritmo *Gradient Boosting*. Este modelo foi capaz de identificar corretamente 76,34% dos casos de fraude e 82,90% de não fraude, com uma precisão geral de 79,62%, indicando ser eficiente no combate às fraudes nos sistemas de abastecimento de água.

Confusion Matrix		
Rows Number : 12002	FRAUDE (Predicted)	NFRAUDE (Predicted)
FRAUDE (Actual)	4581	1420
NFRAUDE (Actual)	1026	4975

Class Statistics									
Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
FRAUDE	4581	1026	4975	1420	76.34%	81.70%	76.34%	82.90%	78.93%
NFRAUDE	4975	1420	4581	1026	82.90%	77.80%	82.90%	76.34%	80.27%

Overall Statistics				
Overall Accuracy	Overall Error	Cohen's kappa ( $\kappa$ )	Correctly Classified	Incorrectly Classified
79.62%	20.38%	0.592	9556	2446

Figura 5: Matriz de confusão do modelo.

Entretanto, o modelo classificou erroneamente 1.420 (23,66%) matrículas como não fraude (falsos negativos) e 1.026 (17,09%) como fraude (falsos positivos), totalizando uma taxa média de erro de 20,38%. No geral, os números mostram que o modelo tem uma alta capacidade de classificar corretamente casos de fraude e não fraude (79,62%), podendo se tornar uma ferramenta importante no combate às ligações clandestinas e perdas aparentes no consumo de água.

A partir da aplicação do modelo desenvolvido, será possível recomendar e direcionar inspeções de forma mais assertiva. Assim, proporcionará ganhos de desempenho para as equipes operacionais, reduzindo custos com inspeções físicas e diminuindo perdas com a regularização dos usuários dos serviços de abastecimento de água.

As fraudes no consumo de água representam uma parte importante da perda de receita das empresas de saneamento e, portanto, devem ser combatidas para evitar a disseminação entre os demais consumidores e a correspondente queda nas receitas da empresa. Por outro lado, a seleção errada de possíveis fraudes pode acabar gerando mais custos com as equipes de inspeção de campo. Por isso, os sistemas de identificação de fraudes devem visar a menor taxa de erro possível.

Para reduzir as taxas de erro do classificador, os resultados devem ser avaliados em busca de padrões entre os dados classificados incorretamente, buscando padrões que possam ajudar na evolução da solução e/ou propostas de novas soluções.

Portanto, comparando os resultados da classificação com os dados originais do conjunto de dados usado para testar o modelo (Conjunto de Teste), temos que o modelo acertou 60% dos consumidores com ligações ativas, 89% dos consumidores com ligações inativas e 67% dos outros tipos. A Figura 6 apresenta a comparação dos resultados por tipo de ligação.

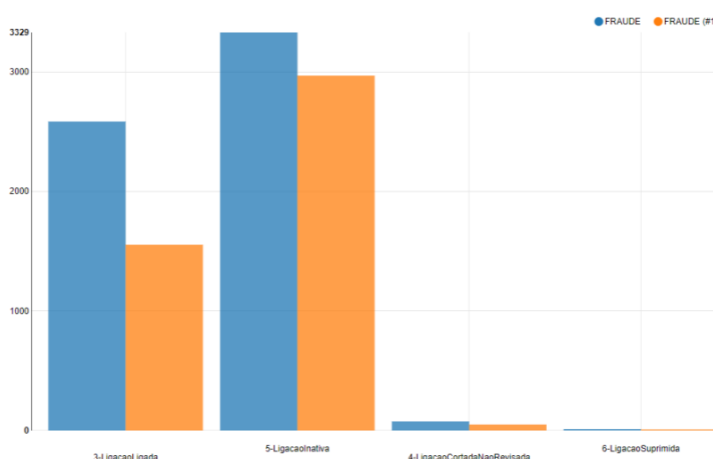


Figura 6: Resultados por tipo de ligação.

Analisando os resultados de falsos negativos (matrículas com fraude classificados como não fraude), pode-se observar que 73% (1.030) são consumidores com conexões ativas, 25% são consumidores com conexões inativas e 2% de outros tipos.

Portanto, embora o modelo apresente resultados promissores no geral, observamos razoável dificuldade do preditor para classificar corretamente as fraudes que ocorreram em consumidores com ligações ativas. Essa complexidade se dá devido às variáveis que influenciam o consumo de água. A redução no consumo pode ser atribuída à existência de algum tipo de fraude ou mudança nos hábitos dos consumidores, redução no número de residentes ou viagens temporárias, por exemplo.

## **CONCLUSÕES E RECOMENDAÇÕES**

Neste trabalho, utilizamos técnicas de aprendizado de máquina para identificar fraudes no consumo de água em uma empresa pública de saneamento. Também foram aplicadas técnicas de pré-processamento de dados, otimização de hiperparâmetros e validação cruzada para construir modelos preditivos baseados em algoritmos de aprendizado supervisionado. Os resultados obtidos demonstram a eficácia do modelo construído. A precisão de 79,62% indica um elevado poder de identificação de casos de fraude e não fraude. O modelo foi treinado com técnicas de rigor científico elevado, contribuindo para o avanço do estado da arte na identificação de fraudes no consumo de água.

Os algoritmos de aprendizado supervisionado funcionam bem a partir do aprendizado com dados disponibilizados para o treinamento do modelo. Para melhorar o desempenho do modelo utilizando algoritmos de classificação, será necessário obter mais exemplos de matrículas com fraudes em ligações ativas e realizar uma análise detalhada para obter variáveis com melhor correlação com as fraudes.

Pesquisas futuras ainda são necessárias para avaliar os resultados obtidos pelos algoritmos de aprendizado de máquina em comparação com a identificação manual por equipes de especialistas. No entanto, considerando a eficácia alcançada pelos modelos desenvolvidos aqui, o potencial para sua incorporação no processo de gestão operacional é claro, podendo auxiliar técnicos na identificação de fraudes no consumo de água, contribuindo para a redução de perdas e aumento da eficiência operacional.

Complementarmente, o desenvolvimento de novos modelos usando outras técnicas de inteligência computacional, como Aprendizagem profunda, Redes Neurais ou Séries Temporais, para uso conjunto ao modelo apresentado. Métodos supervisionados e não supervisionados são ferramentas complementares poderosas para o combate às fraudes, conforme Baensens et al. (2015). De acordo com Phua et al. (2010), o uso de modelos híbridos pode fornecer melhores resultados do que o uso individual de modelos de aprendizado supervisionado. Portanto, sugerimos que mais estudos sejam realizados para melhorar os resultados na identificação de fraudes em conexões ativas.

## REFERÊNCIAS BIBLIOGRÁFICAS

- AESBE (2015). Guia prático para estimação de consumo não autorizados e volumes não apropriados por falhas de cadastro. Disponível em: [http://www.aesbe.org.br/guias\\_praticos/](http://www.aesbe.org.br/guias_praticos/). Acesso em: 16 abr. 2025.
- AL-RADAIDEH, Q. A.; AL-ZOUBI, M. M. A data mining based model for detection of fraudulent behaviour in water consumption. In: INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION SYSTEMS, 9., 2018, Irbid. Proceedings [...]. Irbid: IEEE, 2018. p. 48-54.
- AL-WASHALI, T.; SHARMA, S.; KENNEDY, M. Methods of assessment of water losses in water supply systems: a review. *Water Resources Management*, Dordrecht, v. 30, p. 4985-5001, 2016. DOI: 10.1007/s11269-016-1503-7.
- AL-WASHALI, T.; SHARMA, S.; LUPOJA, R.; AL-NOZAILY, F.; HAIDERA, M.; KENNEDY, M. Assessment of water losses in distribution networks: Methods, applications, uncertainties, and implications in intermittent supply. *Resources, Conservation and Recycling*, Amsterdam, v. 152, p. 104515, 2020. DOI: 10.1016/j.resconrec.2019.104515.
- ANDRADE, R.; SOBRINHO; BORJA, P. C. Gestão das perdas de água e energia em sistema de abastecimento de água da embasa: um estudo dos fatores intervenientes na RMS. *Engenharia Sanitária e Ambiental*, Rio de Janeiro, v. 21, p. 783-795, 2016.
- BAESENS, B.; VAN VLASSELAER, V.; VERBEKE, W. *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. Hoboken: John Wiley & Sons, 2015. 620 p.
- BOLTON, R. J.; HAND, D. J. Statistical fraud detection: A review. *Statistical Science*, v. 17, n. 3, p. 235-249, 2002.
- BRASIL. Sistema nacional de informações sobre saneamento: Diagnóstico dos serviços de água e esgotos - 2017. Brasília, 2018. Disponível em: <http://snis.gov.br/diagnostico-agua-e-esgotos/diagnostico-ae-2017>. Acesso em: 16 abr. 2025.
- BRASIL. Sistema nacional de informações sobre saneamento: Diagnóstico dos serviços de água e esgotos - 2018. Brasília, 2019. Disponível em: <http://www.snis.gov.br/diagnostico-anual-agua-e-esgotos/diagnostico-dos-servicos-de-agua-e-esgotos-2018>. Acesso em: 16 abr. 2025.
- BRASIL. Lei nº 12.026, de 15 de julho de 2020. marco legal do saneamento básico. Diário Oficial da República Federativa do Brasil, Brasília, DF, 2020. Disponível em: <http://www.planalto.gov.br/ccivil/03/Atos/2019-2022/2020/Lei/L14026.html>.
- DE CASTRO FETTERMANN, D.; GUERRA, K. C.; MANO, A. P.; DE ALMEIDA MARODIN, G. Uma sistemática para detecção de fraudes em empresas de abastecimento de água. *Interciencia*, v. 40, p. 114-120, 2015.
- EMBASA. Empresa Baiana de Águas e Saneamento S/A. Disponível em: <https://www.embasa.ba.gov.br/index.php/institucional/a-embasa/apresentacao>. Acesso em: 16 abr. 2025.
- ESPINOSA, F. H. T.; GISSELOT, F. F. P.; ARRIAGADA, I. R. B. Predicción de fraudes en el consumo de agua potable mediante el uso de minería de datos. *Universidad Ciencia y Tecnología*, v. 24, p. 58-66, 2020.
- FAWCETT, T.; PROVOST, F. J. Combining data mining and machine learning for effective user profiling. In: *KDD, 1996, Portland. Proceedings [...]. Portland: KDD, 1996. p. 8-13.*
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, p. 37-37, 1996.
- GOPAL, G. V.; BALAJI, V. Detection of fraudulent behaviour in water consumption using machine learning algorithms. 2020. Trabalho não publicado.

GUMIER, C. C.; LUVIZOTTO JUNIOR, E. et al. Aplicação de modelo de simulação-otimização na gestão de perdas de água em sistemas de abastecimento. Engenharia Sanitária e Ambiental, Rio de Janeiro, 2007.

HAI XIANG, G.; YI JING, L.; SHANG, J.; MINGYUN, G.; YUANYUE, H.; BING, G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, v. 73, p. 220-239, 2017. DOI: <https://doi.org/10.1016/j.eswa.2016.12.035>.

HUDSON, W. D. Reduction of unaccounted-for water. Journal American Water Works Association, v. 56, p. 143-148, 1964.

HUMAI D, E. H.; BARHOUM, T. Water consumption financial fraud detection: A model based on rule induction. In: PALESTINIAN INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY, 2013, Hebron. Proceedings [...]. Hebron: IEEE, 2013. p. 115-120.

IBGE (2020). Instituto Brasileiro de Geografia e Estatística - IBGE. Disponível em: <https://cidades.ibge.gov.br/>. Acesso em: 16 abr. 2025.

IFC (2013). Manual sobre contratos de performance e eficiência para empresas de saneamento em Brasil.

KIM, E.; LEE, J.; SHIN, H.; YANG, H.; CHO, S.; NAM, S.-K.; SONG, Y.; YOON, J.A.; KIM, J.-I. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. Expert Systems with Applications, v. 128, p. 214–224. Disponível em: <https://doi.org/10.1016/j.eswa.2019.03.042>. Acesso em: 16 abr. 2025.

KNIME (2020). Knime analytics platform. Disponível em: <https://www.knime.com/knime-open-source-story>. Acesso em: 16 abr. 2025.

MONEDERO, I.; BISCARRI, F.; GUERRERO, J. I.; ROLDAN, M.; LEON, C. An approach to detection of tampering in water meters. Procedia Computer Science, v. 60, p. 413–421, 2015.

MUTIKANGA, H. E.; SHARMA, S. K.; VAIRAVAMOORTHY, K. Assessment of apparent losses in urban water systems. Water and Environment Journal, v. 25, p. 327–335, 2011. Disponível em: <https://doi.org/10.1111/j.1747-6593.2010.00225.x>. Acesso em: 16 abr. 2025.

MUTIKANGA, H. E.; SHARMA, S. K.; VAIRAVAMOORTHY, K. Methods and tools for managing losses in water distribution systems. Journal of Water Resources Planning and Management, v. 139, p. 166–174, 2013. Disponível em: [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000245](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000245). Acesso em: 16 abr. 2025.

PASSINI, S. R. R.; TOLEDO, C. M. T. Mineração de dados para detecção de fraudes em ligações de água. In: XI SEMINCO-Seminário de computação, 2002.

PHUA, C.; LEE, V.; SMITH, K.; GAYLER, R. A comprehensive survey of data mining-based fraud detection research. arXiv preprint, arXiv:1009.6119, 2010.

QUEIROGA, R. M. Uso de técnicas de data mining para detecção de fraudes em energia elétrica. Biblioteca Central da Universidade Federal do Espírito Santo, 2005.

KUSTERKO, S.; ENSSLIN, S. R.; ENSSLIN, L.; CHAVES, L. C. Gestão de perdas em sistemas de abastecimento de água: uma abordagem construtivista. Engenharia Sanitária e Ambiental, v. 23, 2018.

RASCHKA, S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint, arXiv:1811.12808, 2018.

SAHIN, Y.; BULKAN, S.; DUMAN, E. A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, v. 40, p. 5916–5923, 2013. Disponível em: <https://doi.org/10.1016/j.eswa.2013.05.021>. Acesso em: 16 abr. 2025.

SREEDEVI, E.; SWATHI, M. R. Data mining based model for detection of fraudulent behavior in water consumption, 2021.

SREEKANTH, D.; THINAKARAN, K. Metro water fraudulent prediction in houses using convolutional neural network and recurrent neural network. *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS*, v. 11, p. 1177–1187, 2021.

SRIRAMULU, M.; VANDANA, M.; SWATHI, K.; AMRUTHA, M.; RAHUL, M. Detection of fraudulent behaviour in water consumption, 2020.

TARDELLI FILHO, J. Aspectos relevantes do controle de perdas em sistemas públicos de abastecimento de água. *Revista DAE*, v. 64, p. 6–20, 2016.

THABTAH, F.; HAMMOUD, S.; KAMALOV, F.; GONSALVES, A. Data imbalance in classification: Experimental evaluation. *Information Sciences*, v. 513, p. 429–441, 2020.

TRATA BRASIL, I. Perdas de água 2020 (snis 2018): desafios para disponibilidade hídrica e avanço da eficiência do saneamento básico. São Paulo, 2020. Disponível em:  
[https://www.tratabrasil.org.br/images/estudos/Relatorio\\_Final\\_-\\_Estudo\\_de\\_Perdas\\_2020\\_-\\_JUNHO\\_2020.pdf](https://www.tratabrasil.org.br/images/estudos/Relatorio_Final_-_Estudo_de_Perdas_2020_-_JUNHO_2020.pdf). Acesso em: 16 abr. 2025.

UDDIN, S. A.; AHMED, S. H.; AHMED, F.; RAIS, A. A data mining based model for detection of fraudulent behaviour in water consumption, 2019.

ZYOUD, S. H.; KAUFMANN, L. G.; SHAHEEN, H.; SAMHAN, S.; FUCHS-HANUSCH, D. A framework for water loss management in developing countries under fuzzy environment: Integration of fuzzy ahp with fuzzy topsis. *Expert Systems with Applications*, v. 61, p. 86–105, 2016. Disponível em:  
<https://doi.org/10.1016/j.eswa.2016.05.016>. Acesso em: 16 abr. 2025.