Fraudes no consumo de agua: aplicação de métodos de aprendizado de máquina

Aline Nehls, Andreza Kalbusch, Elisa Henning

INTRODUÇÃO

No setor de abastecimento de água, perdas aparentes são a contabilização de consumo, proveniente de fraude ou erro de medição, sem o devido faturamento (Kusterko et al., 2018). No Brasil, aproximadamente 37,8% da água tratada e distribuída acaba sendo perdida por infrações e vazamentos (SNIS, 2023). Deste modo, este estudo tem como objetivo aplicar métodos de aprendizado de máquina que auxiliem na identificação de fraudes no consumo de água urbano, a fim de prevenir e localizar casos não descobertos.

DESENVOLVIMENTO

O estudo propõe o uso do aprendizado de máquinas a fim de obter predições de fraudes no sistema de distribuição de água. Para realização da pesquisa, utilizou-se uma base de dados contendo o histórico de infrações de cada unidade consumidora de maneira anonimizada. O processamento de dados e implementação dos algoritmos foram realizados por meio de programação em R, com auxílio dos pacotes *readxl, tidyverse, tidymodels, skimr, caret, rpart.plot* e *klaR*.

Os tipos de fraudes analisados foram agrupados em uma única variável categórica, são eles: violação de corte de ramal, ligação direta e *bypass*. Para prever a ocorrência dessas fraudes, foram selecionadas as seguintes variáveis preditoras: Multas anteriores por hidrômetros danificados ou substituídos, multas anteriores por violação de corte de cavalete, quantidade de meses em que se observou queda abrupta no consumo de água, ocorrência de consumo nulo por um período superior a dois meses consecutivos e localização em um distrito de medição de controle (DMC).

O banco de dados forneceu um número muito inferior de fraudes em comparação a não fraudes. Assim, foi necessária a aplicação da técnica de *downsampling* para corrigir o desbalanceamento na fase de treinamento. Em seguida, o conjunto de dados foi dividido de forma a destinar 80% da amostra para treinamento e 20% para o conjunto de testes.

Por fim, empregou-se o aprendizado de máquinas supervisionado para realizar a classificação dos dados de teste. As técnicas aplicadas foram: *Naive Bayes*, árvore de decisão, floresta aleatória, máquina de vetores de suporte e K- vizinhos mais próximos.

RESULTADOS

Após aplicação das técnicas de classificação, os resultados mostraram o modelo de máquina de vetores de suporte (SVM) como o mais consistente na detecção de fraudes. O algoritmo de aprendizagem SVM demonstrou uma maior acurácia na mediana e valores com baixa variação (Tabela 1). Já o modelo com menor variação foi o K-vizinhos mais próximos (KNN). *Naive Bayes* (NB) apresentou alta variabilidade de resultados quando comparado com os outros métodos (figura 1). Os modelos de árvore de decisão (DT) e floresta aleatória (RF) apresentaram resultados ligeiramente inferiores ao SVM, com variabilidade similar.

CONSIDERAÇÕES FINAIS

O estudo possibilitou a análise dos diferentes métodos de aprendizado de máquinas aplicado na predição de fraudes no abastecimento de água com uma base de dados reais. Também, foi possível contornar os desafios causados pela extensão de dados por meio do balanceamento correto com o *downsampling*, permitindo a aplicação dos algoritmos de classificação e evitando distorções no resultado. Assim, demonstrou-se que a implementação de modelos como o SVM podem ser

beneficiais para o combate de infrações na distribuição de água, pois promove a automatização dos processos e diminui a direciona as equipes de fiscalização.

Palavras-chave: aprendizado de máquina; estatística preditiva; detecção de fraudes; perdas aparentes.

ANEXOS

Tabela 1. ***Resultados dos métodos de classificação.***

Método de classificação	Mínimo	Q1	Mediana	Média	Q3	Máximo
Máquina de vetores de suporte (SVM)	0.7914256	0.8365228	0.8498739	0.8198838	0.8233200	0.8253968
Floresta aleatória (RF)	0.7973594	0.8001780	0.8253968	0.8246596	0.8481163	0.8522474
K-vizinhos mais próximos (KNN)	0.8107106	0.8181549	0.8218365	0.8198838	0.8233200	0.8253968
Naive Bayes (NB)	0.5738653	0.8116007	0.8221332	0.7950876	0.8481163	0.9353212
Árvore de decisão (DT)	0.7426580	0.7458834	0.7530040	0.7642266	0.7657618	0.8138258

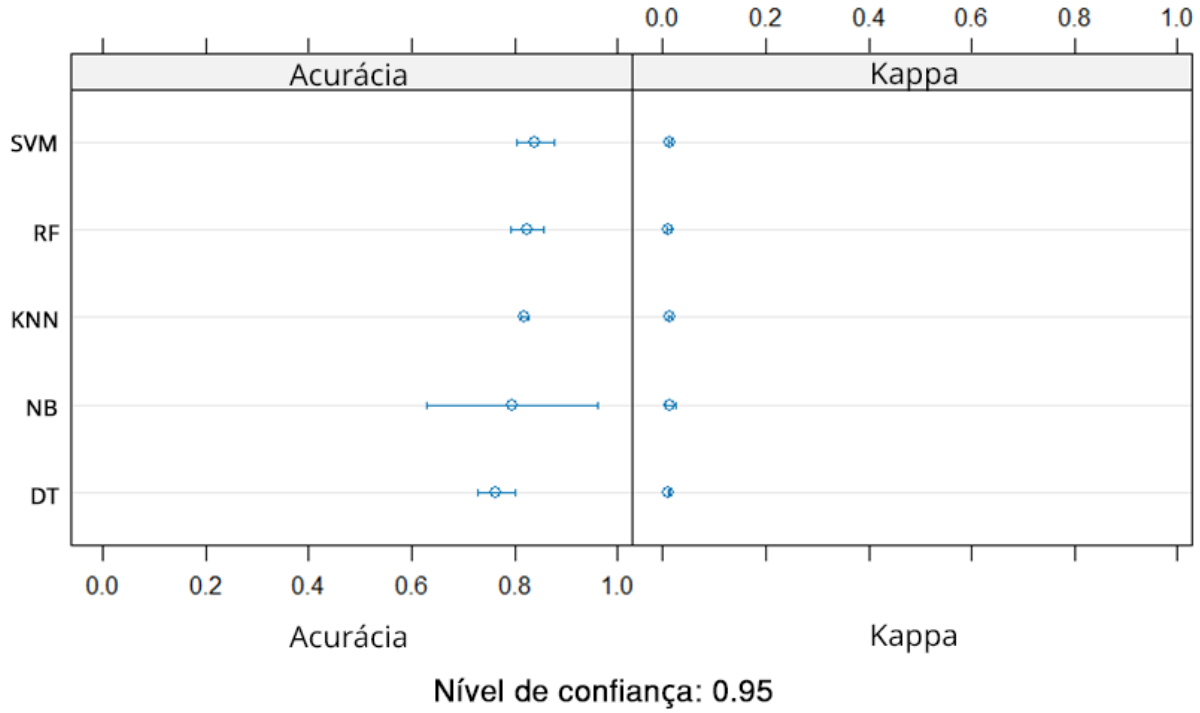


Figura 1. *Gráfico de acurácia e coeficiente Kappa dos métodos utilizados.*

REFERÊNCIAS BIBLIOGRÁFICAS

BRASIL. Ministério das Cidades. Diagnóstico Temático Serviços de Água e Esgoto: Ano de referência 2022. Brasília, DF: SNS, 2023. Disponível em: <https://capacidades.gov.br/capaciteca/diagnostico-tematico-servicos-de-agua-e-esgoto-visao-geral-ano-de-referencia-2022>. Acesso em: 8 set. 2026.

KUSTERKO, Sheila et al. Gestão de perdas em sistemas de abastecimento de água: uma abordagem construtivista. **Engenharia Sanitária e Ambiental**, v. 23, n. 03, p. 615-626, 2018.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Versão 4.6.1. Viena: R Foundation for Statistical Computing, 2026. Disponível em: https://www.R-project.org/.