MÉTODOS DE APRENDIZADO DE MÁQUINA PARA MODELAGEM DE PERDA DE ÁGUA NOS MUNÍCIPIOS BRASILEIROS

João Pedro Felicio Prudencio, Andreza Kalbusch, Daniel Veitex Prates, Elisa Henning

INTRODUÇÃO

Segundo o Sistema Nacional de Informações sobre Saneamento (SNIS), em 2022 aproximadamente 40% da água potável distribuída no Brasil foi perdida (SNIS, 2023), representando desperdício de um recurso vital e agravando problemas ambientais e financeiros. De acordo com Meireles et al. (2023), as perdas de água em sistemas de distribuição se dividem em duas categorias: perda aparente (consumo não autorizado ou erros de medição) e perda real (vazamentos). O estudo de Meireles et al. (2023), realizado em Portugal, apresenta as perdas reais como predominantes, estando associadas a fatores como pressão elevada, infraestrutura envelhecida, extensão da rede, modelo de gestão (público ou privado) e tipologia da área (urbana, rural ou mista). Considerando o impacto desses elementos, pode-se concluir que as perdas de água comprometem a eficiência dos sistemas de distribuição, sendo um desafio para a sustentabilidade hídrica.

No Brasil, o estudo de Gouveia e Soares (2022) utiliza aprendizado de máquina para a previsão de vazamentos de água em tubulações em Brasília, objetivando reduzir perdas de água. Gouveia e Soares (2022) destacam a eficiência de métodos baseados em árvores de decisão, como Floresta Aleatória, *Gradient Boosting* e *XGBoost*. Nesse contexto, o objetivo do presente trabalho é prever a perda de água nos municípios brasileiros, por meio de modelos de aprendizado de máquina baseados em árvores de decisão: Floresta Aleatória, *Gradient Boosting, XGBoost, LightGBM e CatBoost.*

DESENVOLVIMENTO

Foram utilizados dados do ano de 2021 (IBGE, 2021; SNIS, 2021), por ser o ano mais recente que possuía todos os dados desejados para esta pesquisa. As variáveis escolhidas, com base na literatura, incluem: volume de água produzido; volume de água consumido; índice de hidrometração; tarifa média de água; volume de água micromedido; consumo médio per capita de água; quantidade de economias ativas de água; população total do município; extensão da rede de água por ligação; produto interno bruto per capita; extensão da rede de água.

Para a implementação das predições, os dados foram separados aleatoriamente em um conjunto de treino (80%) e teste (20%). Foram implementados então os modelos de aprendizado de máquina: Floresta Aleatória, *Gradient Boosting, XGBoost, LightGBM e CatBoost*. A fim de encontrar a melhor configuração possível para cada modelo, foi realizada a otimização de hiperparâmetros, primeiro com o *GridSearchCV*, que faz pesquisa em grade usando validação cruzada (*Scikit-Learn,* 2025) e depois com o MLJAR, que utiliza otimização bayesiana para encontrar automaticamente as melhores combinações de hiperparâmetros (Płońska; Płoński, 2021). A avaliação do desempenho dos modelos foi feita com quatro métricas: R^2 (Coeficiente de Determinação), RMSE (Erro Quadrático Médio), MAE (Erro Absoluto Médio) e MAPE (Erro Percentual Absoluto Médio).

RESULTADOS

A comparação das métricas dos modelos está na Tabela 1, evidenciando o desempenho superior do *CatBoost*, assim como a Figura 1a), que exibe o gráfico de dispersão com o valor de R^2 do modelo. A importância das variáveis desse modelo, apresentada na Figura 1b), mostrou que

fatores operacionais, especialmente o volume de água produzido e o volume consumido, foram determinantes para a previsão das perdas. Além disso, os aspectos demográficos apresentaram relevância moderada, enquanto os econômicos tiveram contribuição baixa.

CONSIDERAÇÕES FINAIS

Este estudo demonstrou que algoritmos de aprendizado de máquina, com destaque para o *CatBoost*, são eficazes para prever a perda de água em municípios brasileiros. A principal conclusão é que fatores operacionais são determinantes nesse contexto, e, junto aos demográficos, mais relevantes para as perdas do que indicadores puramente econômicos.

Palavras-chave: Perdas de água; Aprendizado de máquina; Regressão linear; Árvore de Decisão.

ILUSTRAÇÕES

Tabela 1. *Métricas de desempenho dos modelos de aprendizado de máquina*

MODELO	R^2	RMSE	MAE	MAPE (%)
Floresta Aleatória	0,514	7,291	5,705	20,11
Gradient Boosting	0,716	5,577	3,844	13,14
XGBoost	0,732	5,412	3,603	12,28
LightGBM	0,713	5,602	3,842	13,14
CatBoost	0,750	5,232	3,383	11,56

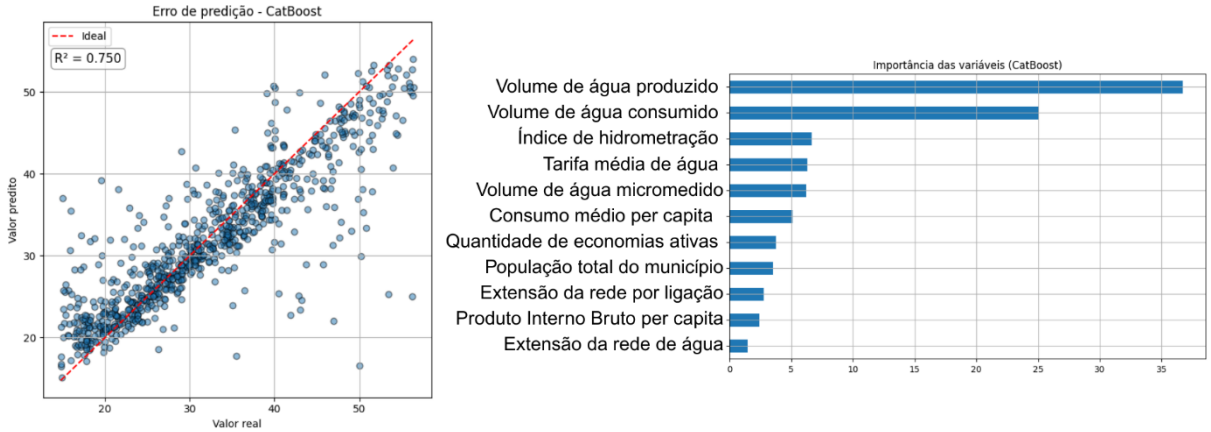

Figura 1. *Gráfico de dispersão entre valores reais e preditos (a) e Importância de variáveis do modelo CatBoost (b)*

REFERÊNCIAS BIBLIOGRÁFICAS

MEIRELES, Inês; SOUSA, Vítor; MATOS, José Pedro; CRUZ, Carlos Oliveira. Determinants of water loss in Portuguese utilities. *Utilities Policy,* v. 83, 2023. https://doi.org/10.1016/j.jup.2023.101603.
GOUVEIA, Cristiano; SOARES, Alexandre. Machine learning classification models applied to water service connection leakage data: contributions on understanding factors involved in failure and insights for infrastructure management. *Environ. Sci. Proc.,* v. 21, 83, 2022. https://doi.org/10.3390/environsciproc2022021083.
SNIS - Série Histórica. Sistema Nacional de Informações de Saneamento, 2021. Disponível em: app4.mdr.gov.br/serieHistorica/municipio/index. Acesso em: 05 set. 2025.
SCIKIT-LEARN. Cross-validation: evaluating estimator performance, 2024. Disponível em: https://scikit-learn.org/stable/modules/cross_validation.html. Acesso em: 10 fev. 2025.
PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825-2830, 2011.
PYTHON SOFTWARE FOUNDATION. *Python Language Reference, version 3.13.3.* Disponível em: http://www.python.org. Acesso em: 15 fev. 2025.
THE PANDAS DEVELOPMENT TEAM. (2020). *pandas-dev/pandas: Pandas. Zenodo*. Disponível em: doi.org/10.5281/zenodo.3509134. Acesso em: 15 fev. 2025.
PŁOŃSKA, Aleksandra; PŁOŃSKI, Piotr. MLJAR: State-of-the-art Automated Machine Learning Framework for Tabular Data. *Version 0.10.3. GitHub repository, 2021*. Disponível em: https://github.com/mljar/mljar-supervised. Acesso em: 15 fev. 2025.
MLJAR. mljar-supervised: Automated Machine Learning Python package for tabular data. 2024. Disponível em: https://supervised.mljar.com. Acesso em: 15 fev. 2025.

DADOS CADASTRAIS

BOLSISTA: João Pedro Felicio Prudencio
MODALIDADE DE BOLSA: PIBIC/CNPq (IC)
VIGÊNCIA: 09/2024 a 08/2025 – Total: 12 meses
ORIENTADOR(A): Elisa Henning
CENTRO DE ENSINO: Centro de Ciências Tecnológicas (CCT)
DEPARTAMENTO: Departamento de Matemática (DMAT)
ÁREAS DE CONHECIMENTO: Engenharias / Engenharia Civil
TÍTULO DO PROJETO DE PESQUISA: Métodos estatísticos e de aprendizado de máquina para análise do consumo de água em edificações.
Nº PROTOCOLO DO PROJETO DE PESQUISA: NPP3195-2025