PERDA DE ÁGUA POTÁVEL NOS MUNICÍPIOS BRASILEIROS: INFLUÊNCIAS DEMOGRÁFICAS, ECONÔMICAS E OPERACIONAIS ANALISADAS POR APRENDIZADO DE MÁQUINA[1]

João Pedro Felicio Prudencio[2], Elisa Henning[3], Andreza Kalbusch[4], Daniel Veitex Prates[5].

1 Vinculado ao projeto "Métodos estatísticos e de aprendizado de máquina para consumo de água em edificações"
[2] Acadêmico do Curso de Bacharelado em Ciência da Computação – CCT – Bolsista PIBIC
[3] Orientadora, Departamento de Matemática – CCT – elisa.henning@udesc.br
[4] Coorientadora, Departamento de Engenharia Civil – CCT – andreza.kalbusch@udesc.br
[5] Mestrando do Programa de Pós-Graduação em Engenharia Civil – CCT

A importância da água para o ser humano é incontestável, seja para o consumo, a agricultura, a higiene, entre outras finalidades, ela se mostra um recurso imprescindível. Nesse contexto, para garantir um fornecimento seguro até a rede doméstica, a água é retirada de fontes naturais, tratada para remover impurezas, armazenada em reservatórios e distribuída por tubulações até finalmente chegar às residências. Entretanto, esse sistema não é perfeito, existem perdas de água durante o transporte. Estima-se que no ano de 2022 no Brasil aproximadamente 40% da água potável distribuída foi perdida, de acordo com o Sistema Nacional de Informações sobre Saneamento (SNIS, 2023). Essa perda representa não apenas um problema financeiro, mas também um desperdício de um recurso vital. Nesse sentido, esta pesquisa faz uso de aprendizado de máquina para analisar o impacto que as variáveis relacionadas ao serviço de fornecimento de água potável possuem na perda de água.

Primeiramente, foi realizada uma revisão da literatura sobre a perda de água no mundo, com o objetivo de aplicar esse conhecimento ao cenário brasileiro. Em seguida, foram coletados dados de 2021 do IBGE (Instituto Brasileiro de Geografia e Estatística) e do SNIS, que contêm informações relevantes para o contexto da perda de água. A base selecionada possui 3996 cidades do Brasil contendo 12 variáveis explicativas, sendo duas do IBGE: PIB e POP_TOT e dez do SNIS: AG002, AG003, AG005, AG006, AG008, AG010, IN005, IN009, IN020, IN022. Além disso, há uma variável dependente (IN049) de perda de água na distribuição. A descrição dessas variáveis pode ser encontrada na Tabela 1.

Em sequência, foi conduzida a análise exploratória dos dados para identificar padrões, tendências e possíveis inconsistências nesses dados. Durante a análise de correlação linear, utilizou-se o coeficiente de correlação de Pearson para avaliar a relação linear entre a variável de perda de água e as demais variáveis. Os resultados indicaram que não há uma correlação forte entre a variável de perda de água e as variáveis explicativas. Além disso, quatro variáveis (PIB, AG003, AG005, IN020_AE) apresentam probabilidade de significância (valor – p) acima de 0,05, ou seja, a relação entre essas variáveis e a variável dependente pode não ser estatisticamente significativa. Os dados foram separados aleatoriamente em um conjunto de treino (80% dos dados) e outro de teste (20% dos dados). Na análise da regressão linear, o coeficiente de determinação (R^2) obtido

com os dados de treino foi de 0,087, o que sugere que o modelo tem uma capacidade limitada de ajuste aos dados de treinamento e explica apenas uma pequena parte da variabilidade nas perdas de água.

Apesar das limitações apontadas por essas métricas, foram implementados dois modelos de aprendizado de máquina para prever as perdas de água. Todos os modelos que serão comentados em sequência foram implementados utilizando a biblioteca Scikit-learn (Pedregosa *et al*., 2011) da linguagem Python. O primeiro modelo utilizado foi a Floresta Aleatória (*Random Forest*), que, segundo a documentação do Scikit-learn, treina várias árvores de decisão em subconjuntos diferentes dos dados originais, combinando suas previsões de forma paralela.

Buscando otimizar a eficiência do modelo de Floresta Aleatória, foram ajustados quatro hiperparâmetros essenciais: o número de árvores na floresta (n_estimators), a profundidade máxima de cada árvore (max_depth), o número mínimo de amostras para dividir um nó (min_samples_split) e o número mínimo de amostras em uma folha (min_samples_leaf). Essa otimização foi realizada utilizando o GridSearchCV, que, de acordo com a documentação do Scikit-learn, testa diferentes combinações desses hiperparâmetros e aplica validação cruzada para encontrar a melhor configuração. A combinação que obteve o melhor coeficiente de determinação foi composta por: *max_depth* de 32, *min_samples_leaf* de 1, *min_samples_split* de 2 e *n_estimators* de 315.

As métricas de desempenho do modelo aplicando o conjunto de teste resultaram em um Erro Quadrático Médio de 7,2919, Erro Absoluto Médio de 5,7053, Coeficiente de Determinação (R^2) de 0,5143 e Erro Percentual Absoluto Médio de 20,11%. Embora a análise de regressão linear inicial tenha indicado um baixo valor de R^2, o modelo de Floresta Aleatória demonstrou uma capacidade de explicação de 51,43% das perdas de água. Ademais, por meio do gráfico de importância de variáveis foi possível observar que fatores relacionados ao serviço de abastecimento e fatores demográficos mostraram maior efeito na melhoria do modelo que fatores econômicos. Além disso, as variáveis que menos tiveram contribuição para o modelo foram justamente as quatro que possuem menor probabilidade de significância.

Subsequentemente, foi realizado o mesmo procedimento de obtenção dos melhores hiperparâmetros, porém dessa vez para o conjunto de dados sem essas quatro variáveis. Embora o GridSearchCV tenha indicado os mesmos valores para os hiperparâmetros, os resultados foram diferentes. As métricas obtidas foram: Erro Quadrático Médio de 6,662, Erro Absoluto Médio de 4,96, Coeficiente de Determinação (R^2) de 0,594 e Erro Percentual Absoluto Médio de 17,43%.

O segundo modelo implementado foi o *Gradient Boosting*, uma variação da floresta aleatória. Conforme a documentação do Scikit-learn, o Gradient Boosting funciona a partir de um processo de "*boosting*", onde as árvores de decisão são treinadas sequencialmente, com cada nova árvore corrigindo os erros das anteriores, aprimorando gradualmente a predição do conjunto de dados. Esse aprimoramento é controlado por uma taxa de aprendizado que regula a contribuição de cada árvore na sequência.

Para aumentar a eficiência desse modelo, foram otimizados cinco hiperparâmetros: os mesmos quatro utilizados na Floresta Aleatória (*max_depth, min_samples_leaf, min_samples_split e n_estimators*), além de um novo hiperparâmetro, o *learning_rate*, que representa a taxa de aprendizado. O modelo foi implementado com a mesma separação de treino e teste usada na Floresta Aleatória, com todas as 12 variáveis. Novamente, o processo de otimização foi realizado com o GridSearchCV, resultando em uma configuração otimizada com *max_depth* de 4, *min_samples_leaf* de 2, *min_samples_split* de 5, *n_estimators* de 750 e *learning_rate* de 0,12. As

métricas de desempenho do modelo, quando aplicadas ao conjunto de teste, mostraram um Erro Quadrático Médio de 5,577, Erro Absoluto Médio de 3,844, Coeficiente de Determinação (R^2) de 0,716 e Erro Percentual Absoluto Médio de 13,14%. Ou seja, comparado à Floresta Aleatória, o Gradient Boosting não apenas apresentou erros menores, como também mostrou capacidade de explicação de 71,6% das perdas de água. O resultado do cálculo da importância das variáveis apresentou resultados semelhantes ao modelo anterior.

Nesse sentido, o procedimento anterior foi realizado, aplicando a otimização de hiperparâmetros para o conjunto de dados com essas variáveis removidas. A configuração otimizada apresenta um *max_depth* de 4, *min_samples_leaf* de 2, *min_samples_split* de 4, *n_estimators* de 800 e *learning_rate* de 0,1. Os resultados apresentaram uma leve melhoria, com um Erro Quadrático Médio de 5,385, Erro Absoluto Médio de 3,478, Erro Percentual Absoluto Médio de 11,93% e Coeficiente de Determinação (R^2) de 0,735. O comparativo das métricas de desempenho dos quatro modelos está na Tabela 2. O gráfico comparando os valores reais e previstos pelo modelo *Gradient Boosting* com as variáveis selecionadas está na Figura 1.

Em suma, a análise dos modelos desenvolvidos demonstrou que fatores demográficos, como a população total do município, e aspectos diretamente ligados ao abastecimento de água exercem influência nas perdas de água potável. Em contrapartida, fatores econômicos mostraram-se menos relevantes nesse cenário específico. Assim, a aplicação de modelos de aprendizado de máquina provou ser uma abordagem promissora para investigar as variáveis que impactam as perdas de água potável nos municípios brasileiros.

Tabela 1. *Variáveis usadas nos modelos de aprendizado de máquina.*

CÓDIGO	DESCRIÇÃO
IN049	Percentual de perdas na distribuição
PIB	Produto Interno Bruto per capita do município
POP_TOT	População total do município
AG002	Quantidade de ligações ativas de água, providas ou não de hidrômetro
AG003	Quantidade de economias ativas de água
AG005	Extensão de rede de água
AG006	Volume de água produzido
AG008	Volume de água micromedido
AG010	Volume de água consumido
IN005	Tarifa média de água
IN009	Índice de hidrometração
IN020	Extensão da rede de água por ligação
IN022	Consumo médio per capita de água

Tabela 2. *Métricas de desempenho dos modelos de aprendizado de máquina.*

MODELO	R^2	RMSE	MAE	MAPE
Floresta Aleatória com todas as variáveis	0,514	7,291	5,705	20,11
Floresta Aleatória apenas com as variáveis selecionadas	0,594	6,662	4,960	17,43
***Gradient Boosting* com todas as variáveis**	0,716	5,577	3,844	13,14
***Gradient Boosting* apenas com as variáveis selecionadas**	0,735	5,385	3,478	11,93

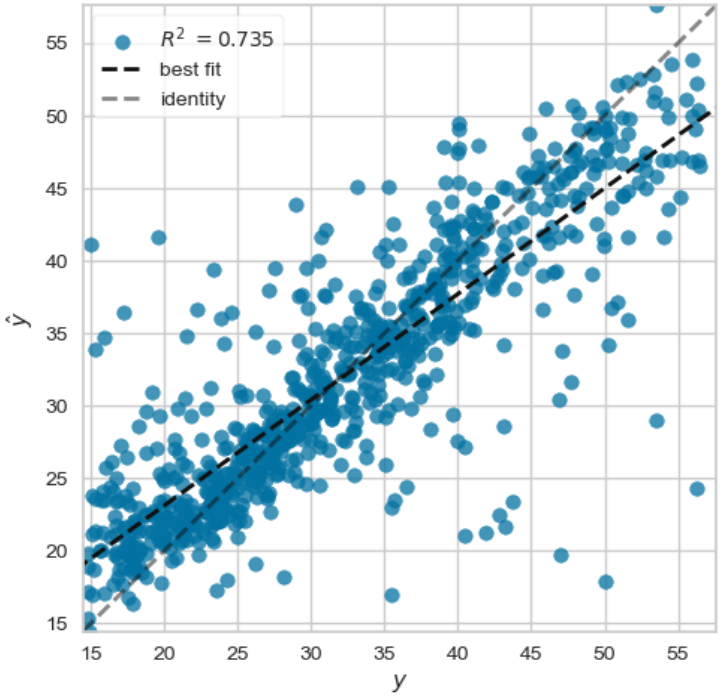

Figura 1. *Gráfico de dispersão entre valores reais e preditos pelo modelo Gradient Boosting*

Referências

PEDREGOSA, F. *et al*. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

SNIS - Série Histórica. Sistema Nacional de Informações de Saneamento, 2021. Disponível em: app4.mdr.gov.br/serieHistorica/municipio/index. Acesso em: 25 jun. de 2024.

SNIS. Abastecimento de Água - 2022. Disponível em: www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/saneamento/snis/painel/ab. Acesso em: 07 jul. 2024.

Palavras-chave: Perda de água; Aprendizado de máquina; Regressão linear.