

APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA PARA ANÁLISE E PREVISÃO DE CONSUMO DE ÁGUA EM MUNICÍPIOS

Isabela das Chagas Luiz, Andreza Kalbusch, Elisa Henning

INTRODUÇÃO

Este trabalho tem como objetivo investigar como características climáticas e geográficas, aspectos sociais, demográficos e econômicos afetam a demanda de água. A questão central da pesquisa é entender e prever o consumo de água em cidades brasileiras. A identificação desses fatores aprofunda a compreensão do fenômeno, possibilita identificar oportunidades de otimização do consumo e contribui para previsões mais precisas. Esta pesquisa parte da hipótese de que é possível modelar o consumo de água com muitos preditores e alto grau de colinearidade entre eles. Diante disso, foram utilizadas as informações de 1746 municípios brasileiros, sendo aplicados e comparados os modelos de Regressão Linear, Árvore de Decisão e Floresta Aleatória, seguidos pelos modelos de regressão Ridge, Lasso e Elastic Net.

DESENVOLVIMENTO

A base de dados utilizada provém do trabalho de Veiga (2020) e as variáveis utilizadas englobam aspectos climáticos e geográficos (temperatura, precipitação, altitude), demográficos (população, urbanização, densidade, moradores por domicílio, gênero), econômicos (Produto Interno Bruto, renda, preço da tarifa de água), relativos ao fornecimento de água (hidrometração) e sociais (uso de máquina de lavar, Índice de Desenvolvimento Humano - IDH, acesso à água canalizada), oferecendo uma visão multidimensional sobre fatores que influenciam o consumo de água. A análise exploratória estatística descritiva foi realizada obtendo informações como média, desvio-padrão e histogramas de cada variável.

A base de dados foi dividida em 80% para treino e 20% para teste, sendo realizada também a validação cruzada. Foram aplicados os modelos de Regressão Linear, Árvore de Decisão e Floresta Aleatória, com previsões de consumo de água obtidas a partir da influência das variáveis. Posteriormente, foram realizados os modelos de Regressão Ridge, Lasso e Elastic Net.

Para a aplicação da Regressão Linear, as variáveis crianças, adolescentes e IDH foram retiradas, devido ao valor obtido do Fator de Inflação da Variância (VIF), indicando correlação entre essas variáveis. Conforme Kutner *et al.* (2005), essa retirada é necessária apenas em casos de alta multicolinearidade. Nos modelos Ridge, Lasso e Elastic Net, os coeficientes permanecem estabilizados mesmo na presença de variáveis correlacionadas (Quantum Technologies, 2024).

As métricas utilizadas para avaliação dos modelos incluíram erro quadrático médio (RMSE), coeficiente de determinação (R^2) e erro médio absoluto (MAE). A importância relativa das variáveis foi visualizada por meio de gráficos, permitindo identificar quais preditores possuem maior impacto no consumo de água. As análises foram conduzidas no software R (R Core Team, 2025), utilizando o pacote Caret (Kuhn, 2008), com nível de significância adotado de 5%.

RESULTADOS

Durante a análise exploratória apuramos que a média de consumo nas cidades analisadas foi de 122 litros/habitantes/dia com desvio-padrão de 32,6 litros/habitantes/dia. A partir das métricas dos modelos aplicados a Floresta Aleatória apresentou o melhor desempenho, destacando-se pelos menores valores de erro e maior capacidade preditiva. As previsões são

apresentadas na Figura 1, em que a Figura 1a) exibe a importância relativa das variáveis no modelo, evidenciando que o preço da tarifa de água foi o preditor de maior relevância. Na Figura 1b) é mostrado o gráfico de predições do consumo de água (litros/habitantes/dia), no qual se observa uma relação linear entre os valores estimados e os valores reais, indicando que o modelo tem potencial na tarefa de previsão. A análise desses gráficos é importante, pois mostra a confiabilidade do modelo em reproduzir o padrão observado nos dados e fornece subsídios para ajustar estratégias de gestão do consumo de água.

CONSIDERAÇÕES FINAIS

O trabalho investigou como características climáticas e geográficas, aspectos sociais, demográficos e econômicos afetam a demanda de água, e os resultados também evidenciaram o potencial preditivo dos modelos aplicados. Para continuidade da pesquisa, pretende-se realizar a modelagem para os municípios separados por porte, bem como a aplicação de outros métodos, como Mínimos Quadrados Parciais.

Palavras-chave: consumo de água; análise de regressão; aprendizado de máquina.

ILUSTRAÇÕES

Tabela 1. Métricas de desempenho dos modelos aplicados.

Modelo	RMSE	R ²	MAE
Regressão Linear	24,22	0,45	17,15
Árvore de Decisão	27,58	0,31	19,78
Floresta Aleatória	18,71	0,69	12,62
Regressão Ridge	20,62	0,53	14,98
Regressão Lasso	20,50	0,53	15,01
Elastic Net	20,50	0,53	15,02

Fonte: Autor, 2025.

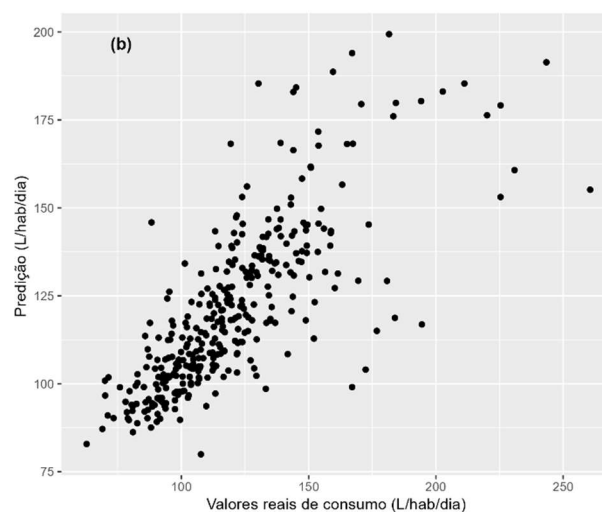
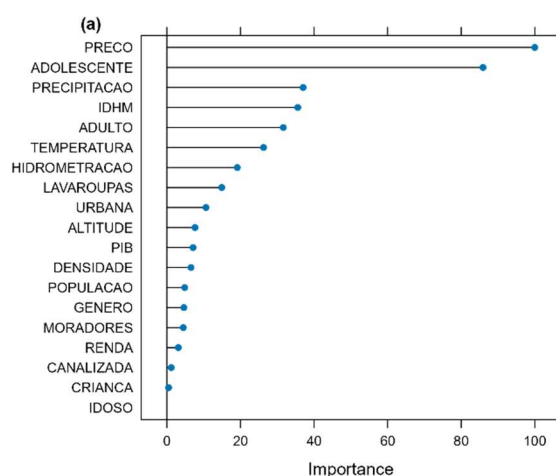


Figura 1. Resultados do melhor dos modelos realizados. (a) importância das variáveis para o modelo Floresta Aleatória, (b) predição do modelo Floresta Aleatória. Fonte: Autor, 2025.

REFERÊNCIAS BIBLIOGRÁFICAS

CUANTUM TECHNOLOGIES. **Feature Engineering for Modern Machine Learning with Scikit-Learn**. 1. ed. 2024. Disponível em: <https://www.cuantum.tech/books/feature-engineering-for-modern-machine-learning-with-scikitlearn?srsId=AfmBOoqIjZG7kfp36x5nzz7grCII3Rgx-S4rdQkz1JjG9XpVRo0fYzdp>. Acesso em: 28 ago. 2025.

KUHN, Max. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, n. 5, p. 1–26, 2008. Disponível em: <https://doi.org/10.18637/jss.v028.i05>. Acesso em: 28 ago. 2025.

KUTNER, Michael H.; NACHTSHEIM, Christopher J.; NETER, John; LI, William. **Applied Linear Statistical Models**. 5. ed. Boston: McGraw-Hill, 2005.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2025.

VEIGA, Fabiano da. **Análise da influência dos fatores climáticos, geográficos, demográficos, econômicos e sociais no consumo de água no Brasil**. Dissertação (Mestrado Acadêmico em Engenharia Civil) – Programa de Pós-Graduação em Engenharia Civil, Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Florianópolis, 2020. Disponível em: <https://pergamumweb.udesc.br/acervo/150155>. Acesso em: 28 ago. 2025.

DADOS CADASTRAIS

BOLSISTA: Isabela das Chagas Luiz

MODALIDADE DE BOLSA: PIBIC/CNPq (IC)

VIGÊNCIA: 09/24 a 08/25 – Total: 12 meses

ORIENTADOR(A): Elisa Henning

CENTRO DE ENSINO: CCT

DEPARTAMENTO: Departamento de Matemática

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