MODELAGEM DO CONSUMO DE ÁGUA EM EDIFÍCIOS DE ESCRITÓRIO ATRAVÉS DE MÉTODOS DE MACHINE LEARNING[1]

Ana Luíza Bruhn[2], Andreza Kalbusch[3], Elisa Henning[4].

[1] Vinculado ao projeto "Investigação de fatores relacionados ao consumo de água no ambiente construído"
[2] Acadêmica do Curso de Engenharia Civil – CCT – Bolsista PIBIC/CNPq
[3] Orientadora, Departamento de Engenharia Civil – CCT – andreza.kalbusch@udesc.br
[4] Coorientadora, Departamento de Engenharia Civil – CCT – elisa.henning@udesc.br

A escassez de água tem se tornado uma questão relevante na atualidade, com a mudança dos efeitos climáticos e estilo de vida da população. Assim, torna-se necessário o estudo e análise de fatores que influenciam o consumo de água, o que pode ser útil para elaboração de políticas públicas para uso racional da água. Para a previsão de consumo de água, têm-se utilizado diversos métodos e abordagens estatísticas. Os desafios em relação à previsão da demanda de água consistem na decisão de quais variáveis e dados devem ser utilizados; e qual técnica de modelagem é a mais indicada para cada cenário, a fim de obter elevada acurácia preditiva (Lee; Derrible, 2020). Nesse contexto, há lacunas de estudos relacionados ao consumo de água na tipologia de edifícios de escritórios. Portanto, o objetivo desta pesquisa consiste na análise e emprego de diferentes modelos estatísticos: regressão linear múltipla, regressão por penalização *ridge*, *lasso*, *elastic net* e *stacked regression*, para decisão do melhor algoritmo a ser empregado na previsão do consumo de água em escritórios.

Em relação à utilização de modelos de regressão, em que se busca relacionar variáveis preditoras a uma variável resposta, é comum a ocorrência da multicolinearidade. Na análise da regressão linear múltipla, a multicolinearidade indica a relação linear entre as variáveis independentes, que pode levar a resultados distorcidos em relação à compreensão dos fatores que mais influenciam na variável resposta (Shrestha, 2020). Técnicas de regressão por regularização, como *ridge*, *lasso* e *elastic net*, objetivam minimizar as consequências da multicolinearidade entre as variáveis do banco de dados evitando o ajuste excessivo do modelo aos dados de treino (*overfitting*) e aumentando a interpretabilidade do modelo com um nível de acurácia maior (Thevaraja; Rahman; Gabirial, 2019).

Já a *stacked regression* é um método que combina diferentes modelos com o objetivo de atingir maior precisão preditiva. Nessa abordagem, as previsões dos modelos criadas no nível anterior são consideradas regressoras de entrada para modelos do próximo nível. O método busca minimizar o erro quadrático médio da validação cruzada do modelo obtido. A Figura 1 apresenta a lógica da *stacked regression* deste estudo e revela os diferentes modelos empregados. Assim, os modelos base para a construção da combinação foram regressão linear múltipla, regressão por penalização *ridge*, *lasso* e *elastic net*. Depois, utilizando as previsões criadas pelos modelos base, o meta modelo selecionado foi a regressão linear múltipla.

Objetivando a avaliação dos modelos, três métricas foram utilizadas: erro absoluto médio (MAE), erro quadrático médio (RMSE) e coeficiente de determinação (R^2). A Tabela 1 apresenta os resultados obtidos para cada modelo na base de dados de treino, em reamostragem e validação cruzada de 10 vezes. Devido à proximidade das medidas de erro (MAE e RMSE) para todos os

modelos, selecionou-se o modelo combinado *stacked-lm* como o melhor em acurácia preditiva, devido ao seu maior coeficiente de determinação (0,63).

Em relação às limitações deste estudo, destaca-se o tamanho da amostra (n=52) que dificultou a obtenção de resultados ainda mais precisos da modelagem estatística. Ressalta-se que foram entrevistados os responsáveis por 165 escritórios, no entanto as demais edificações não puderam compor a amostra final por não apresentarem características essenciais para realização desta pesquisa, como a medição individualizada do consumo de água. Esta pesquisa, entretanto, serve de guia para casos com pequenas amostras e pode auxiliar na decisão do melhor método estatístico a ser empregado para previsão do consumo de água em edifícios de escritório. Portanto, como sugestão para trabalhos futuros, pode-se estender a pesquisa e coleta de dados a outras localidades, a fim de aumentar a amostra. Também sugere-se usar diferentes técnicas de reamostragem para a etapa de validação dos modelos.

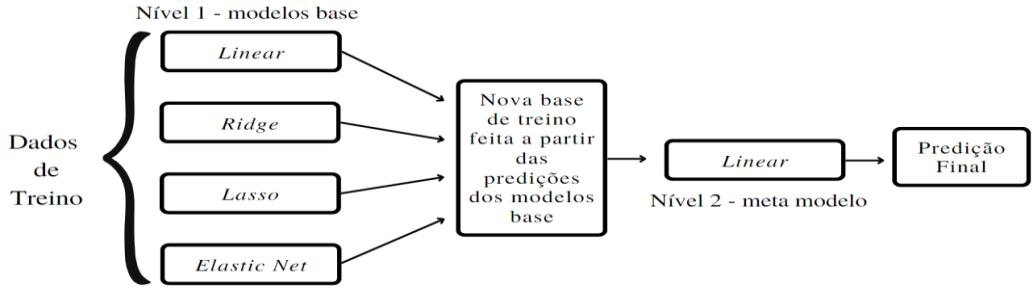

Figura 1. *Lógica da stacked regression.*

Tabela 1. *Resultados obtidos das métricas de avaliação para a base de dados de treino.*

Modelo	MAE	RMSE	R^2
Ridge	35,1	41,9	0,46
Lasso	31,4	38,2	0,49
Elastic Net	30,9	37,8	0,45
Stacked-lm	32,9	38,1	0,63

REFERÊNCIAS

LEE, Dongwoo; DERRIBLE, Sybil. Predicting residential water demand with machine-based statisticallearning. **Journal of Water Resources Planning and Management**, v. 146, n. 1, 2020.

SHRESTHA, Noora. Detecting multicollinearity in regression analysis. **American Journal of Applied Mathematics and Statistics**, v. 8, n. 2, p. 39-42, 2020.

THEVARAJA, Mayooran; RAHMAN, Azizur; GABIRIAL, Mathew. Recent developments in data science: Comparing linear, ridge and lasso regressions techniques using wine data. In: **International Conference on Digital Image and Signal Processing 2019: DISP 2019**. University of Oxford, 2019. p. 1-6.

Palavras-chave: Consumo de água. Escritórios. *Machine Learning*.