PREVISÃO DE NÍVEL HIDROLÓGICO NA BACIA DO JAGUARÃO COM XGBOOST E ANÁLISE SHAP DE IMPORTÂNCIA DE VARIÁVEIS

Alejandra Bustillos Vega[1]; André Rodrigues Ferreira[2]

Palavras-Chave – XGBoost; Random Forest; SHAP; previsão hidrológica; bacia do Jaguarão.

INTRODUÇÃO

No Rio Grande do Sul, inundações causam perdas humanas e econômicas em escala crescente: o evento de maio de 2024 destruiu 9.300 residências e gerou prejuízos estimados em R$ 87 bilhões, com quase 200 óbitos (Collischonn et al., 2025; Marengo et al., 2024). A bacia do Jaguarão, na fronteira com o Uruguai, apresenta histórico recorrente de cheias que atingem Jaguarão e Arroio Grande. Sistemas de previsão com antecedência de 1 a 2 dias permite acionar protocolos de evacuação antes que a cota crítica seja atingida.

Modelos de aprendizado de máquina (ML), em especial XGBoost e Random Forest, têm superado abordagens hidrológicas clássicas em previsões de séries temporais (Chen e Guestrin, 2016). O obstáculo central é que esses modelos funcionam como caixas-pretas, limitando sua aceitação operacional por gestores de risco (Zounemat-Kermani e Kheimi, 2026). Métodos de Inteligência Artificial Explicável (XAI), em especial os valores SHAP (Lundberg e Lee, 2017), permitem verificar se o modelo aprendeu relações fisicamente plausíveis, conectando desempenho estatístico à interpretação hidrológica (Wei et al., 2024; Bezerra et al., 2026). Este trabalho treina e avalia modelos XGBoost e Random Forest para prever a cota diária na estação 88300000 (rio Jaguarão) e aplica análise SHAP para validar a consistência física dos modelos.

DADOS E METODOLOGIA

Foram consolidadas séries horárias de dados hidrométricos e meteorológicos da bacia do Jaguarão (2007–2024), provenientes das redes INMET (A827, A836), ANA(88300000) e CEMADEN (430130501A). A variável-alvo é a cota diária na estação 88300000. O preenchimento de falhas foi realizado com KNN (k=5), toda a série original foi preservada. A engenharia de features produziu 37 variáveis preditoras em cinco grupos: (i) memória hidrológica, lags de 1, 2, 7, 14, 30 dias e médias móveis de 7, 14, 30 dias; (ii) tendência, delta de cota em 1 e 7 dias; (iii) precipitação acumulada em janelas de 1, 3, 7, 30 e 90 dias; (iv) dados da estação upstream 88260000, vazão e cota com lags de 1–3 dias; (v) variáveis temporais cíclicas. A análise abrange dois períodos independentes: 2017–2021 (1.657 registros) e 2015–2020 (2.006 registros).

Os modelos XGBoost (n=200, max_depth=7, lr=0,1) e Random Forest (n=200, max_depth=15) foram avaliados com divisão temporal estrita 80%/20%: os primeiros 80% em ordem cronológica constituem o treino e os últimos 20% o teste, eliminando vazamento de informação futura. O StandardScaler foi ajustado exclusivamente nos dados de treino. A validação cruzada temporal usou TimeSeriesSplit com 5 dobras. A interpretabilidade foi avaliada por importância por ganho (XGBoost), importância por permutação (Random Forest) e valores SHAP (Molnar, 2022).

RESULTADOS

Os modelos apresentaram desempenho elevado nos dois períodos (Figura 1; Tabela 1). Para 2017–2021, o XGBoost registrou R^2=0,9937, RMSE=4,05 cm, MAE=1,74 cm e MAPE=1,57%; o Random Forest obteve R^2=0,9909, RMSE=4,86 cm e MAPE=2,03%. Para 2015–2020, o XGBoost alcançou R^2=0,9968, RMSE=3,69 cm e MAPE=2,08%, com o Random Forest em R^2=0,9874, RMSE=7,31 cm e MAPE=2,98%. O CV R^2 temporal (TimeSeriesSplit, 5 dobras) variou de 0,942 a 0,982, confirmando que os resultados não são produto de overfitting. O XGBoost superou consistentemente o Random Forest, com erros absolutos abaixo de 2,12 cm em ambos os períodos.

Período	Modelo	R^2	RMSE (cm)	MAE (cm)	MAPE (%)
2017–2021	XGBoost	0,9937	4,05	1,74	1,57
	Random Forest	0,9909	4,86	2,28	2,03
2015–2020	XGBoost	0,9968	3,69	2,11	2,08
	Random Forest	0,9874	7,31	3,03	2,98

Tabela 1 – Desempenho dos modelos XGBoost e Random Forest (split temporal 80/20, sem remoção de outliers). Fonte: elaboração própria.

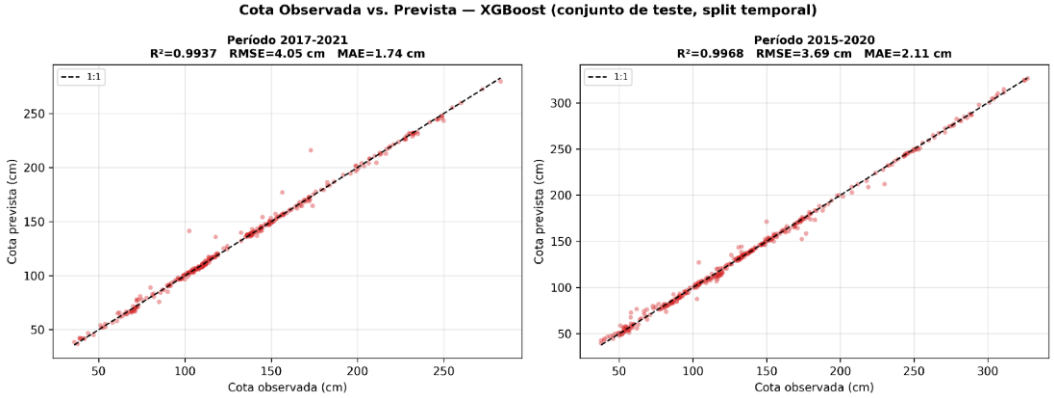


Figura 1. Dispersão cota observada vs. cota prevista pelo XGBoost (conjunto de teste, últimos 20% em ordem cronológica) para os períodos 2017–2021 e 2015–2020. Linha tracejada: concordância perfeita (1:1). Fonte: elaboração própria.

A análise de importância de features revelou padrões fisicamente coerentes (Figura 2). Em 2017–2021, cota_lag_1 concentra a maior fração da importância por ganho, com contribuições secundárias de cota_ma_7 e delta_cota_1d. Em 2015–2020, emerge co-dominância entre cota_lag_1 e cota_ma_7, a média móvel de 7 dias ganha peso equivalente em períodos com maior diversidade hidrológica. A importância por permutação (Random Forest) confirma cota_lag_1 como preditor dominante em ambos os períodos.

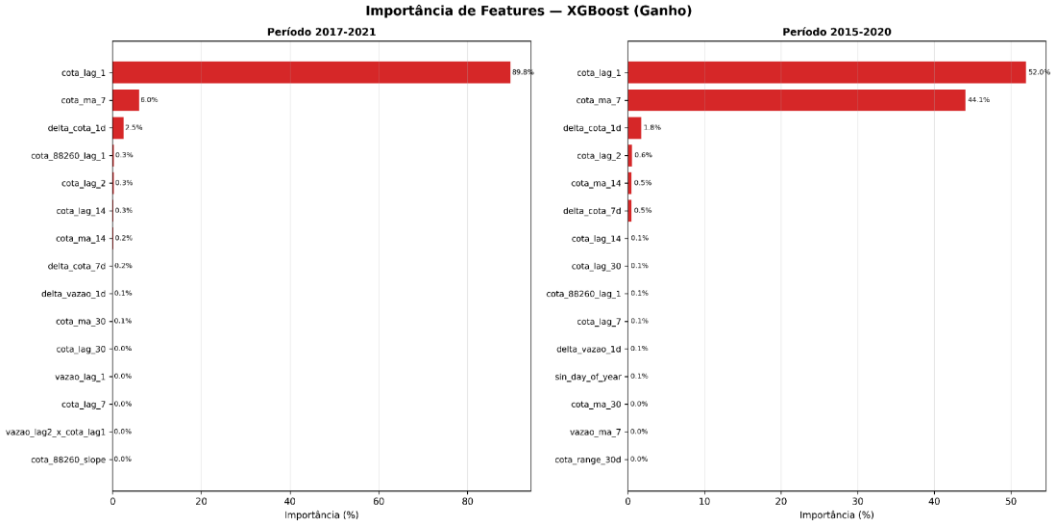


Figura 2. Importância de features pelo método de ganho (XGBoost) para os períodos 2017–2021 (esq.) e 2015–2020 (dir.). Em 2017–2021, cota_lag_1 domina; em 2015–2020, co-dominância com cota_ma_7. Fonte: elaboração própria.

A análise SHAP (Figura 3) confirma a interpretação física. Cota_lag_1 alto (vermelho) gera contribuições SHAP positivas expressivas, o modelo aprendeu a inércia hidrológica. O

delta_cota_1d positivo eleva a previsão (rio em subida → mais provável de continuar subindo); a cota_ma_7 funciona como âncora de médio prazo. A cota_88260_lag_2, nível upstream com 2 dias de atraso, aparece com importância positiva, capturando o tempo de viagem hidráulico entre as estações (~14,4 km, 48 h). Esse alinhamento entre estatística e física viabiliza o uso operacional do modelo em gestão de risco (Bezerra et al., 2026).

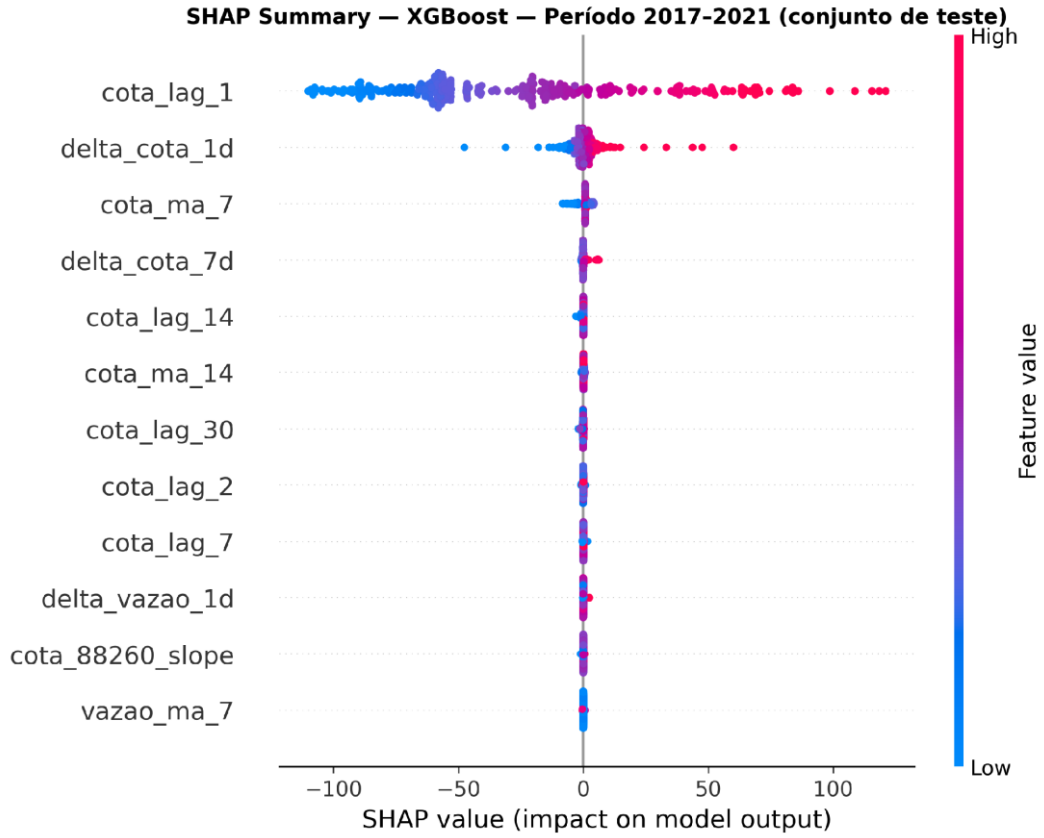


Figura 3. SHAP Summary Plot — XGBoost, período 2017–2021 (conjunto de teste, split temporal). Posição horizontal: magnitude e direção da contribuição; cor: valor da feature (vermelho=alto, azul=baixo). Fonte: elaboração própria.

CONCLUSÕES

O XGBoost superou o Random Forest em ambos os períodos, com R^2>0,993, RMSE entre 3,69 e 4,05 cm e MAPE<2,1%. Resultados obtidos com dados completos e validados com split temporal estrito, garantindo ausência de vazamento de informação e confiabilidade para uso operacional.

A feature cota_lag_1 foi o preditor dominante, confirmando a forte autocorrelação do nível do Jaguarão. Em 2015–2020, cota_ma_7 atingiu importância equivalente, mostrando que a tendência de médio prazo é crítica em horizontes de maior variabilidade.

Os valores SHAP revelaram relações físicas plausíveis: cota_lag_1 alto → previsão alta; delta_cota_1d positivo → tendência de alta; cota_88260_lag_2 → tempo de viagem hidráulico de 48 h capturado autonomamente pelo modelo. Esse alinhamento entre desempenho e física valida a interpretabilidade do modelo para uso em sistemas de alerta operacionais (Bezerra et al., 2026). Trabalhos futuros incluirão extensão para as bacias do Camaquã e Guaíba e incorporação de precipitação por ensemble.

REFERÊNCIAS

BEZERRA, R.; ELEUTÉRIO, J.; SOLHA, P.; BRENTAN, B.; DE MELLO, C.; HERRERA, M.; RODRIGUES, A. (2026). "Towards more accurate very short-range flood forecasts with physics-enhanced machine learning models." Journal of Hydroinformatics, v.28, n.1.

CHEN, T.; GUESTRIN, C. (2016). "XGBoost: A Scalable Tree Boosting System." In: Proceedings of the 22nd ACM SIGKDD. New York: ACM, pp. 785–794.

COLLISCHONN, W. et al. (2025). "The May 2024 flood disaster in Rio Grande do Sul, Brazil: hydrological analysis and reflections on future preparedness." Hydrological Sciences Journal.

LUNDBERG, S. M.; LEE, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems, v.30, pp. 4765–4774.

MARENGO, J. A. et al. (2024). "Eventos extremos e inundações no Brasil: impactos e perspectivas." Revista Brasileira de Meteorologia.

MOLNAR, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2ª ed. Disponível em: christophm.github.io/interpretable-ml-book.

WEI, Z. et al. (2024). "Explainable Artificial Intelligence for flood forecasting using deep learning: A review." Journal of Hydrology.

ZOUNEMAT-KERMANI, M.; KHEIMI, M. (2026). "Explainable AI in hydrology and hydrogeology: state-of-the-art and future prospects." Journal of Hydroinformatics.

AGRADECIMENTOS

Os autores agradecem à FAPERGS pelo financiamento do projeto de pós-doutorado e às agências ANA, INMET e CEMADEN pelo fornecimento dos dados utilizados neste estudo.