

XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

CLASSIFICAÇÃO DE REGIME DE VAZÕES USANDO APRENDIZADO DE MÁQUINA: ESTUDO DE CASO NO AÇUDE ORÓS-CE

Tatiane Lima Batista¹; Larissa Zaira Rafael Rolim²; José Kerlly Soares de Araújo³; Jorgiane Pires Bezerra⁴, Alan Michell Barros Alexandre⁵ & Francisco de Assis de Souza Filho⁶

Abstract: Understanding the streamflow regime is important for water resources management, especially in areas with hydrological and climatic vulnerabilities such as the semi-arid region of Ceará, Brazil. This study explored the use of machine learning methods for classifying the inflow regime of a strategic reservoir located in the state of Ceará, based on two climate predictors: the Atlantic Dipole and Niño3 index. The performances of four methods were compared: Decision Tree (DT), Support Vector Machine (SVM), Multinomial Logistic Regression (MLR), and k-Nearest Neighbors (k-NN), aiming to classify the years as Dry, Normal, or Wet. The k-NN method showed the best overall classification performance, while the Decision Tree model yielded the worst results. In general, the models showed higher accuracy in identifying dry years. On the other hand, the Normal category had the lowest accuracy rates, indicating greater difficulty for the models in distinguishing years with average flow regimes. Some models classified only a few years as wet, showing a certain degree of conservatism for this category. Future studies should investigate approaches based on multi-model or hybrid models to achieve a more robust representation of the hydrological regime in the region.

Resumo: O conhecimento do regime de vazões é importante no contexto da gestão de recursos hídricos em áreas com vulnerabilidades hídricas e climáticas como o semiárido cearense. Este trabalho abordou o uso de métodos de aprendizado de máquina para classificação do regime de vazões afluentes a um reservatório estratégico localizado no estado do Ceará a partir de dois preditores climáticos: Dipolo do Atlântico e Nino3. Foram comparados os desempenhos de quatro métodos: árvore de decisão (DT), máquina de vetores de suporte (SVM), regressão logística multinomial (MLR) e k-vizinhos mais próximos (k-NN) para classificar os anos como Secos, Normais ou Chuvosos. O método k-NN apresentou o melhor desempenho geral na classificação, enquanto o modelo baseado em Árvore de Decisão (DT) obteve os piores resultados. De maneira geral, os modelos apresentaram maior capacidade de acerto na identificação dos anos secos. Por outro lado, a categoria Normal foi a que obteve os menores índices de acerto, indicando maior dificuldade dos

1) Universidade Federal do Ceará – Departamento de Engenharia Hidráulica e Ambiental, Campus do Pici – Bloco 713. Cep: 60400-900. Fortaleza – Ceará, Brasil. E-mail: tatiane@crateus.ufc.br

2) Universidade Federal do Ceará – Departamento de Engenharia Hidráulica e Ambiental, Campus do Pici – Bloco 713. Cep: 60400-900. Fortaleza – Ceará, Brasil. E-mail: larissazairarr@gmail.com

3) Universidade Federal do Ceará – Departamento de Engenharia Hidráulica e Ambiental, Campus do Pici – Bloco 713. Cep: 60400-900. Fortaleza – Ceará, Brasil. E-mail: josekerlly@alu.ufc.br

4) Universidade Federal do Ceará – Departamento de Engenharia Hidráulica e Ambiental, Campus do Pici – Bloco 713. Cep: 60400-900. Fortaleza – Ceará, Brasil. E-mail: jorgianepires@alu.ufc.br

5) Universidade Federal do Ceará, Campus de Crateús - Avenida Professora Machadinho Lima, S/N, Príncipe Imperial, Crateús-CE, CEP 63708-825. Fone: +55 (88) 3691-9700 - e-mail: alanmichell@crateus.ufc.br

6) Universidade Federal do Ceará – Departamento de Engenharia Hidráulica e Ambiental, Campus do Pici – Bloco 713. Cep: 60400-900. Fortaleza – Ceará, Brasil. E-mail: assis@deha.ufc.br

modelos em distinguir anos com regime médio de vazões. Alguns modelos classificaram poucos anos como chuvosos, apresentando certo conservadorismo para essa categoria. Sugere-se a investigação de abordagens baseadas em multimodelos ou modelos híbridos, buscando uma representação mais robusta do regime hidrológico na região.

Palavras-Chave – Temperatura da Superfície do Mar, Árvore de Decisão, Máquina de Vetor de Suporte, k-NN, Regressão Logística.

INTRODUÇÃO

A dinâmica hidrológica do semiárido brasileiro, particularmente no estado do Ceará, é fortemente modulada por fenômenos climáticos interanuais de origem oceânica. A concentração da precipitação nos meses de fevereiro a maio, aliada à alta variabilidade interanual e aos baixos índices pluviométricos médios, torna o sistema de reservatórios do semiárido cearense sensível às variações da temperatura da superfície do mar (TSM) nos oceanos Pacífico e Atlântico Tropical. Entre as principais forçantes oceânicas de larga escala destacam-se o índice Niño3, associado ao fenômeno El Niño–Oscilação Sul (ENOS), e o Dipolo do Atlântico (Pinheiro; Rocha; Drumond, 2021).

A atuação combinada de anomalias positivas no Pacífico Equatorial Central (El Niño) e de uma configuração positiva do Dipolo do Atlântico está associada ao deslocamento meridional da Zona de Convergência Intertropical (ZCIT), com impactos negativos sobre a precipitação no Nordeste brasileiro (Furtado *et al.*, 2025). No caso do Ceará, esse padrão climático reduz a afluência dos principais reservatórios estaduais, como o açude Orós, o que afeta diretamente a alocação de água e o planejamento de operação dos sistemas hídricos (Martins *et al.*, 2018).

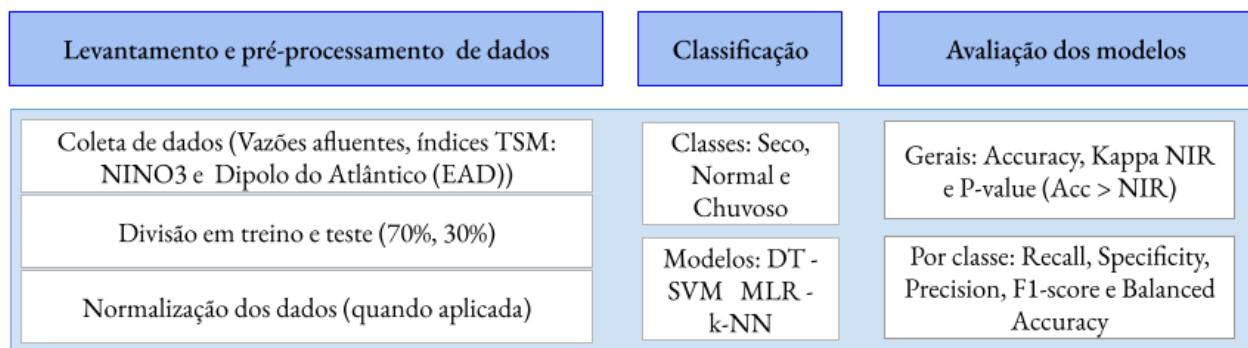
A previsibilidade sazonal de variáveis hidrológicas como a vazão tem sido explorada a partir de modelos estatísticos e de aprendizado de máquina, com destaque para abordagens que utilizam médias sazonais de índices oceânicos como variáveis explicativas. Estratégias baseadas em técnicas como redes neurais, análise de componentes principais e classificação supervisionada têm apresentado bons resultados na estimativa de variáveis hidrológicas em bacias do Nordeste, especialmente quando associadas ao monitoramento de TSM (Cardoso; Malfatti; Dias, 2022; Toledo *et al.*, 2023).

Neste estudo, investiga-se a aplicação de algoritmos de classificação supervisionada para prever o regime hidrológico anual do açude Orós com base em preditores oceânicos. A variável de resposta é a classe de vazão anual (Seco, Normal ou Chuvoso), definida a partir dos tercís da série histórica (1951–2021). Como variáveis explicativas, utilizam-se as médias trimestrais dos índices Niño3 e Dipolo do Atlântico (OND do ano t–1). São comparados os desempenhos de quatro métodos: árvore de decisão (DT), máquina de vetores de suporte (SVM), regressão logística multinomial (MLR) e k-vizinhos mais próximos (k-NN).

MÉTODO

A pesquisa segue três etapas principais (Figura 1), iniciando com a preparação do conjunto de dados, seguindo pela aplicação dos modelos de classificação e com a análise dos resultados a partir de métricas de avaliação. A área de estudo será a bacia hidrográfica do açude Orós, localizado na Região Hidrográfica do Alto Jaguaribe. O açude tem relevância estratégica no contexto dos recursos hídricos do estado, sendo o maior da região e atuando também na perenização do trecho a jusante do Rio Jaguaribe (COGERH, 2021).

Figura 1 – Delineamento da pesquisa



Levantamento e pré-processamento dos dados

Os dados utilizados foram: (1) Label: Classificação dos anos em “Seco”, “Normal” ou “Chuvoso” a partir das vazões afluentes ao açude Orós: média de Janeiro-Dezembro do ano t. (1951 a 2021), (2) Índices climáticos: NINO3 e Dipolo do Atlântico (EAD) – foram utilizadas as médias trimestrais de outubro-novembro-dezembro (OND) do ano anterior (t-1). (1950 a 2020). Os dados dos preditores climáticos foram obtidos no endereço eletrônico do *International Research Institute for Climate and Society* (IRI) e do portal do *Earth System Research Laboratory* (ESRL). Cada ano foi classificado em “Seco”, “Normal” ou “Chuvoso” a partir do valor da vazão afluente anual média obtida do Projeto Alocar (COGERH, 2021). Os tercis da série de vazões foram usados para separar a série nas três categorias.

Os dados foram divididos em dois conjuntos: um de treino (1951 a 2000) e um de teste (2001 a 2021). A normalização dos dados foi realizada para os modelos SVM, MLR e k-NN. Foi aplicada padronização (*z-score normalization*) que, para cada variável, subtrai a média e divide pelo desvio padrão. Foram aplicados os mesmos parâmetros de padronização do treino nos dados de teste. O modelo DT não necessita de normalização porque suas decisões não são tomadas em cálculos de distância ou magnitude absoluta. As árvores são invariantes a escalas diferentes entre os atributos.

Modelos de classificação

Árvore de Decisão (Decision Tree, DT)

As árvores de decisão (DTs) são métodos amplamente utilizados em aprendizagem supervisionada, representando decisões e possíveis desfechos em uma estrutura semelhante a um fluxograma, o que facilita sua interpretação (Breiman, 1984). A construção da árvore envolve divisões sucessivas dos dados com base nos valores das variáveis de entrada; cada nó interno avalia uma variável, enquanto cada folha representa uma previsão (Ragettli *et al.*, 2017). As divisões geralmente buscam minimizar critérios como variância ou erro quadrático médio. Para prever novos casos, percorre-se a árvore desde a raiz até uma folha, tomando decisões conforme os valores das variáveis de entrada. Para melhorar o desempenho e evitar sobreajuste, é comum realizar a poda da árvore (Xu e Liang, 2021). Entre as principais vantagens desse método está a interpretabilidade, porém as árvores de decisão apresentam instabilidade estatística: pequenas mudanças nos dados de treinamento podem gerar árvores com estruturas bastante diferentes, mesmo após a poda (Hastie *et al.*, 2009).

Máquina de Vetor de Suporte (Support Vector Machine, SVM)

As Máquinas de Vetores de Suporte (SVM) são algoritmos de aprendizado de máquina reconhecidos pela boa capacidade de generalização, alcançada ao minimizar o risco estrutural sobre um conjunto limitado de padrões de aprendizagem (Baydaroglu e Koçak, 2014). O objetivo do SVM é encontrar um hiperplano que separe os pontos de diferentes classes, ou, em problemas de regressão, prediga a variável de interesse, sempre buscando maximizar a margem entre o hiperplano e os pontos mais próximos, chamados vetores de suporte (Pham *et al.*, 2016).

O SVM pode lidar com dados não linearmente separáveis por meio de funções kernel, que mapeiam os dados para um espaço de maior dimensão onde a separação linear se torna possível (Tehrany *et al.*, 2014). O hiperplano ótimo é calculado a partir de uma função de decisão baseada nos vetores de suporte, suas classes e a função kernel escolhida, geralmente do tipo linear, polinomial, sigmoide ou base radial (RBF) (Wang *et al.*, 2022). Estudos mostram que o kernel RBF é especialmente eficaz para problemas hidrometeorológicos (Leong *et al.*, 2021), sendo sua fórmula dada pela Equação 1.

$$K(x_i, x_j) = \exp(-\gamma x_i - x_j)^2 \quad (1)$$

Em que: γ controla o grau de não linearidade do modelo. Valores muito baixos ou altos de γ podem levar a underfitting ou overfitting, respectivamente.

K-vizinhos mais próximos (k- nearest neighbors, k-NN)

O método dos k-vizinhos mais próximos (kNN) é um algoritmo simples de aprendizado supervisionado, aplicável a problemas de classificação e regressão (Cover; Hart, 1967). Na classificação, para um novo ponto de classe desconhecida, o algoritmo identifica os k exemplos mais próximos e, atribui a classe mais frequente entre os vizinhos. (Silva *et al.*, 2021). O parâmetro k controla o viés e a variabilidade do modelo: valores pequenos aumentam o risco de overfitting e valores grandes podem levar a underfitting (James *et al.*, 2021). Para este método foi aplicada a validação cruzada para encontrar o valor do parâmetro k.

Regressão logística multinomial (Multinomial logistic regression, MLR)

A Regressão Logística Multinomial é uma extensão da regressão logística binária para problemas com três ou mais categorias (Hosmer *et al.*, 2013). O modelo estima simultaneamente a probabilidade de cada classe usando um modelo linear para cada uma, comparada a uma categoria de referência. As probabilidades são obtidas por meio da função softmax (Kleinbaum; Klein, 2010).

Avaliação dos modelos

Na etapa de avaliação foram aplicadas métricas de avaliação geral do desempenho de cada modelo e métricas de avaliação por categoria frequentemente utilizadas na literatura em problemas de classificação.

As métricas de avaliação geral foram: Acurácia (Accuracy), o Coeficiente Kappa de Cohen (Kappa), a Taxa de Informação Nula (No Information Rate – NIR) e o valor-p (P-value) associado ao teste da hipótese de que a acurácia do modelo é significativamente superior à NIR. O Kappa avalia o nível de concordância entre as classificações previstas pelo modelo e as observações reais, ajustando a acurácia para o caso de acertos por acaso, sendo considerado mais robusto que a simples acurácia em situações com distribuição desigual entre as classes. A NIR representa a taxa de acertos que seria alcançada simplesmente escolhendo a classe mais frequente em todo o conjunto de dados, funcionando como uma linha de base para comparação. O valor-p associado ao teste da NIR indica se a acurácia observada é estatisticamente superior ao desempenho esperado por uma classificação aleatória baseada na NIR (Kuhn, Johnson, 2013)

As métricas de avaliação específicas por categoria foram: Sensibilidade (Recall), Especificidade (Specificity), Precisão (Precision), Medida-F1 (F1-score) e Acurácia Balanceada (Balanced Accuracy).

A Accuracy expressa a proporção de acertos em relação ao total de casos analisados, sendo uma medida geral de desempenho do modelo. A Specificity mede a capacidade do modelo em identificar corretamente os casos negativos, ou seja, os casos que não pertencem à classe em questão. A Precision indica a fração de casos classificados como positivos que, de fato, são positivos. A Recall representa a capacidade do modelo em identificar corretamente os casos positivos reais. Já o F1-score corresponde à média harmônica entre a Precision e a Recall. Já a Balanced Accuracy calcula a média entre a Sensibilidade e a Especificidade. (Grus, 2016).

RESULTADOS E DISCUSSÕES

A análise dos resultados inicia-se com a Tabela 1, a qual apresenta a comparação entre as classes observadas e as classes previstas por cada modelo de classificação (SVM, DT, MLR e KNN) para o período de 2001 a 2021. As classes categóricas (Seco, Normal e Chuvoso) representam diferentes regimes de vazão, sendo utilizadas como alvo para os modelos preditivos.

Tabela 1 – Classes Observadas e Previsões por Modelos de Classificação

Ano	Observado	SVM	DT	MLR	KNN
2001	Seco	Normal	Chuvoso	Chuvoso	Normal
2002	Normal	Seco	Seco	Seco	Seco
2003	Normal	Normal	Normal	Seco	Chuvoso
2004	Chuvoso	Normal	Normal	Seco	Chuvoso
2005	Seco	Seco	Normal	Seco	Seco
2006	Normal	Seco	Seco	Seco	Seco
2007	Normal	Seco	Seco	Seco	Normal
2008	Chuvoso	Normal	Normal	Normal	Normal
2009	Chuvoso	Normal	Normal	Normal	Normal
2010	Seco	Normal	Normal	Seco	Normal
2011	Chuvoso	Chuvoso	Seco	Normal	Chuvoso
2012	Seco	Chuvoso	Seco	Normal	Chuvoso
2013	Seco	Seco	Seco	Seco	Seco
2014	Seco	Seco	Seco	Seco	Seco
2015	Seco	Seco	Seco	Seco	Seco
2016	Seco	Seco	Normal	Seco	Seco
2017	Seco	Normal	Normal	Normal	Normal
2018	Normal	Normal	Normal	Normal	Normal
2019	Normal	Chuvoso	Chuvoso	Chuvoso	Chuvoso
2020	Chuvoso	Chuvoso	Chuvoso	Chuvoso	Chuvoso
2021	Normal	Chuvoso	Seco	Normal	Normal

Observa-se que modelos como o SVM e o KNN tendem a apresentar maior estabilidade nas previsões para anos classificados como “Seco”, enquanto os anos classificados como “Chuvoso”

apresentam maior variabilidade entre os modelos. Além disso, é possível perceber que os modelos apresentam diferentes níveis de sensibilidade para identificar os anos “Normais”, que frequentemente são confundidos com “Seco” ou “Chuvoso”, refletindo a maior ambiguidade dessa categoria intermediária.

A avaliação desses resultados é apresentada em duas etapas. Primeiramente, comparou-se o desempenho geral dos modelos por meio de medidas globais, cujos valores são sumarizados na Tabela 2. Esses indicadores fornecem uma visão ampla da capacidade preditiva de cada algoritmo, considerando o acerto total e o grau de concordância com a classificação observada, corrigido pelo acaso.

Tabela 2 – Estatísticas de Desempenho Global dos Modelos de Classificação

Modelo/Métrica	Accuracy	Kappa	NIR	P-value (Acc > NIR)
SVM	0.4286	0.0935	0.4286	0.5829
DT	0.3333	-0.0426	0.4286	0.8655
MLR	0.4286	0.0935	0.4286	0.5829
KNN	0.5238	0.2808	0.4286	0.2527

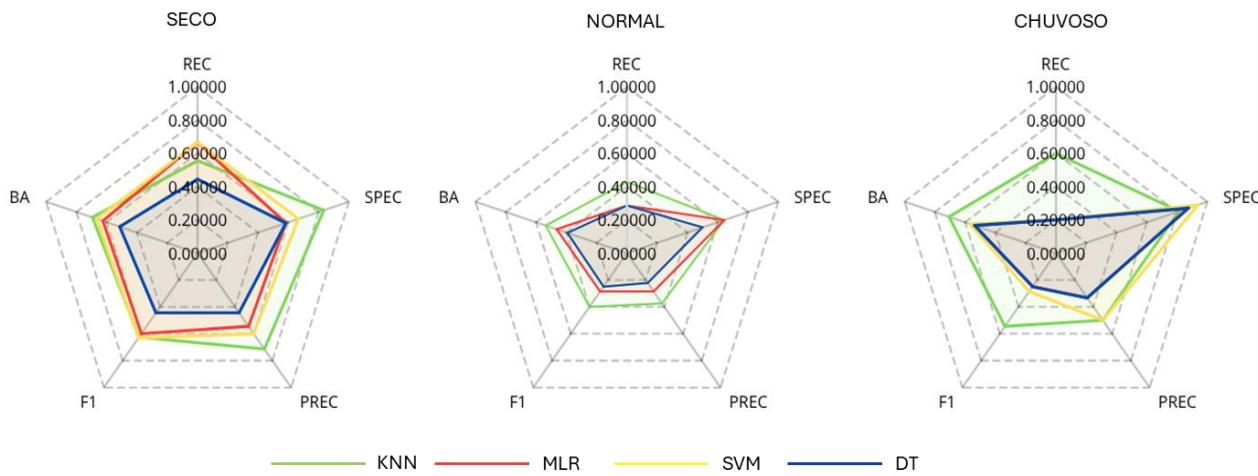
Em seguida, realizou-se uma análise mais detalhada, segmentada por classe, com o objetivo de identificar quais categorias de ano (Seco, Normal e Chuvoso) são melhor representadas por cada modelo. Essa abordagem permite uma avaliação mais precisa dos pontos fortes e limitações individuais dos classificadores, especialmente considerando que, embora o conjunto de treino apresente uma distribuição relativamente balanceada entre as classes, o conjunto de teste possui leve predominância da classe “Seco” e sub-representação da classe “Chuvoso”. Tal assimetria pode influenciar os resultados globais, favorecendo modelos com maior sensibilidade à classe mais frequente, o que reforça a importância de uma análise discriminada por categoria.

A Tabela 2 apresenta os resultados globais dos modelos SVM, DT, MLR e KNN com base nas métricas de acurácia, índice kappa, NIR (No Information Rate) e o p-valor do teste de hipótese para avaliar se a acurácia é significativamente superior ao NIR. Observa-se que o modelo KNN obteve o melhor desempenho global, com acurácia de 0,52 e índice kappa de 0,28, indicando uma concordância moderada com as observações. Embora o p-valor (0,25) não seja estatisticamente significativo ao nível de 5%, ele ainda sugere um desempenho melhor que o aleatório.

Os modelos SVM e MLR apresentaram acurácia de 0,42, exatamente igual ao NIR, com índices kappa modestos (0,09), indicando fraca concordância além do acaso. Os p-valores elevados (> 0.5) reforçam a ausência de evidência estatística de desempenho superior ao classificador aleatório. O modelo DT teve o pior desempenho, com acurácia de apenas 0,33 e índice kappa negativo (-0,04), sugerindo que suas classificações não só não superaram o acaso, mas apresentaram tendência de erro.

De modo geral, os resultados mostram que, apesar de pequenas variações nas acuráncias, o KNN foi o único modelo com desempenho visivelmente superior ao NIR, ainda que sem significância estatística. Já os demais modelos apresentaram limitações claras, com baixa capacidade discriminativa nas classes e desempenho similar ou inferior a um classificador aleatório.

Figura 2 – Métricas de Desempenho por Classe dos Modelos de Classificação



Siglas: REC – Recall; PREC – Precision; F1 – F1-score; SPEC – Specificity; BA – Balanced Accuracy.

A Figura 2 apresenta a avaliação do desempenho dos modelos de classificação considerando as métricas Recall, Specificity, Precision, F1-score e Balanced Accuracy para cada uma das três categorias apresentadas.

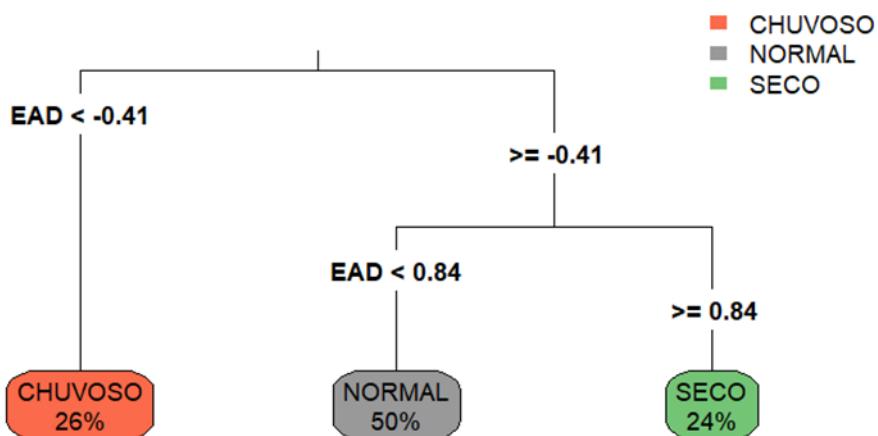
Na classe “Seco”, observou-se que os modelos apresentaram, em geral, desempenho mais satisfatório em comparação às demais classes. O modelo SVM obteve o maior valor de F1-score (0,6316), revelando um bom equilíbrio entre precisão e sensibilidade. Além disso, alcançou um recall de 0,6667, igual ao obtido pelo modelo MLR, indicando boa capacidade de detecção dos anos secos. Por outro lado, o KNN destacou-se pela maior especificidade (0,83) e precisão (0,71), sugerindo maior acurácia ao evitar falsos positivos nessa categoria. O modelo DT apresentou desempenho inferior, com Balanced Accuracy de apenas 0,51, inferior ao observado nos demais modelos.

Em relação à classe “Normal”, todos os modelos demonstraram baixo desempenho. Os valores de recall foram inferiores a 0,43, o que indica dificuldade em identificar corretamente os anos classificados como normais. O KNN, ainda assim, alcançou o melhor F1-score (0,4), embora a Balanced Accuracy tenha sido limitada a 0,53. Os modelos SVM e DT apresentaram os menores valores de Balanced Accuracy (0,39), revelando baixa efetividade na classificação desta categoria.

Na classe “Chuvoso”, os modelos exibiram desempenho intermediário. O KNN novamente se destacou, obtendo os maiores valores de recall (0,6), F1-score (0,54) e Balanced Accuracy (0,70), indicando maior sensibilidade e equilíbrio na classificação dos anos chuvosos. O modelo SVM, embora tenha apresentado elevada especificidade (0,93), evidenciou recall reduzido (0,2), o que sugere uma tendência conservadora, classificando poucos anos como chuvosos. Tanto o MLR quanto o DT apresentaram resultados similares e modestos, com F1-scores de 0,25 e Balanced Accuracy de 0,53.

Além disso, a visualização da árvore de decisão (Figura 3) evidencia que o modelo ignorou o índice NINO3, utilizando apenas a variável EAD como critério de divisão. Isso indica que, para este conjunto de dados e parametrização, o NINO3 não contribuiu de maneira significativa para a classificação dos regimes de vazão, reforçando a importância da análise de importância das variáveis e da engenharia de atributos no desempenho de modelos baseados em árvore.

Figura 3 - Árvore de decisão



Em síntese, o modelo KNN apresentou desempenho mais consistente e equilibrado entre as classes, destacando-se principalmente nas categorias “Seco” e “Chuvoso”. O SVM obteve bom desempenho para “Seco”, mas não se mostrou eficaz nas demais. Os modelos MLR e DT apresentaram desempenho mais limitado, especialmente na classe “Normal”, sendo o DT o menos eficaz entre os avaliados. Pela abertura do gráfico aranha, é possível notar o destaque para o modelo k-NN nas três categorias (linha verde na Figura 2). Também é possível notar que as aberturas são menores na categoria Normal, indicando piores resultados, e maiores na categoria Seco, indicando melhores resultados.

CONCLUSÕES

Este trabalho avaliou quatro métodos de aprendizado de máquina para o problema de classificação do regime de vazões afluentes ao açude Orós, localizado no semiárido nordestino a partir de preditores climáticos associados à Temperatura de Superfície do Mar.

As métricas de avaliação para o grupo de teste, de forma geral, não representaram valores muito altos. A média das acurárias dos quatro modelos foi de cerca de 43% de acertos, sendo o k-NN o modelo que mais acertou, com cerca de 52% de acertos e o *Decision Tree* o modelo que menos acertou, com cerca de 33% de acertos. Apenas o k-NN apresentou valor de acurácia superior ao NIR, ou seja, a taxa de acertos que seria alcançada escolhendo a classe mais frequente em todo o conjunto de dados.

Em conclusão, a análise comparativa dos modelos de classificação evidenciou que o k-NN foi o método que apresentou melhor desempenho na classificação dos regimes de vazão no reservatório estudado, apresentando bons resultados nas categorias “Seco” e “Chuvoso”. O SVM apresentou melhor desempenho na detecção dos anos secos, porém com desempenho limitado nas demais classes, principalmente devido à sua tendência conservadora na classificação dos anos chuvosos. Já os modelos MLR e DT apresentaram resultados mais modestos, com destaque negativo para a baixa

capacidade de classificação da categoria “Normal”, e o DT sendo o menos eficaz de todos. Esses resultados indicam que, apesar dos avanços obtidos, ainda há desafios na correta classificação dos anos com regime médio de vazão, sugerindo a necessidade de estudos futuros que explorem abordagens multimodelos ou híbridas para aprimorar a previsão hidrológica na região.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da coordenação de aperfeiçoamento de pessoal de nível superior – Brasil (CAPES) – código de financiamento 001

REFERÊNCIAS

- BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C. *Classification and regression trees*. Boca Raton: CRC Press, 1984.
- CARDOSO, A. O.; MALFATTI, M. G. L.; DIAS, P. L. S. Predictability of extreme streamflow in Brazilian rivers through climate indices. *International Journal of Hydrology*, v. 6, p. 213-222, 2022.
- COGERH – COMPANHIA DE GESTÃO DOS RECURSOS HÍDRICOS. *Plano de Recursos Hídricos da Região Hidrográfica do Alto Jaguaribe*. Fortaleza: COGERH, 2022.
- COGERH. *Relatório de cálculo das afluências aos reservatórios estratégicos do Ceará: definição das vazões oficiais*. Projeto ALOCAR, 2021.
- COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21-27, 1967.
- FURTADO, T. M. S. et al. Variability of the Atlantic Niño: Impacts on precipitation in the state of Maranhão, Brazil. *Atmosfera*, v. 39, 2025.
- GRUS, J. *Data Science do Zero*. São Paulo: Starlin Alta Editora e Consultoria Eireli, 2016.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. *The elements of statistical learning: data mining, inference, and prediction*. 2. ed. New York: Springer, 2009.
- HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. 3. ed. New York: Wiley, 2013.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning: with Applications in R*. 2. ed. New York: Springer, 2021.
- KLEINBAUM, D. G.; KLEIN, M. *Logistic Regression: A Self-Learning Text*. 3. ed. New York: Springer, 2010.
- KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. New York: Springer, 2013.
- MARTINS, E. S. P. R. et al. A multimethod attribution analysis of the prolonged Northeast Brazil hydrometeorological drought (2012–16). In: HERRERA, S.; HERRERA, C. M. (ed.). *Explaining Extreme Events of 2016 from a Climate Perspective. Bulletin of the American Meteorological Society*, v. 99, n. 1, p. S66–S69, 2018. DOI: 10.1175/BAMS-D-17-0102.1.
- PINHEIRO, E.; ROCHA, R. P.; DRUMOND, A. Assessment of 20th-century reanalysis circulation patterns associated with El Niño–Southern Oscillation impacts on the tropical Atlantic and

northeastern Brazil rainy season. *International Journal of Climatology*, v. 41, n. 7, p. 3824-3840, 2021.

RAGETTLI, S. et al. Modeling flash floods in ungauged mountain catchments of China: A decision tree learning approach for parameter regionalization. *Journal of Hydrology*, v. 555, p. 330-346, 2017. DOI: <https://doi.org/10.1016/j.jhydrol.2017.10.031>.

SILVA, L. A.; PERES, S. M.; BOSCAROLI, C. *Introdução à Mineração de dados: com aplicações em R*. Rio de Janeiro: LTC, 2021.

TOLEDO, J. F. et al. Impacto dos índices climáticos na previsão de séries mensais de vazões. *IEEE Access*, v. 11, p. 21451-21464, 2023.

XU, T.; LIANG, F. Machine learning for hydrologic sciences: An introductory overview. *Wiley Interdisciplinary Reviews: Water*, v. 8, n. 5, p. e1533, 2021. DOI: <https://doi.org/10.1002/wat2.1533>.