

Previsão Antecipada de Dias com Chuva Forte em Parauapebas-PA utilizando XGBoost com Otimização Bayesiana

Angello Gabriel Alves Gonçalves¹

¹Faculdade de Engenharia Elétrica – Universidade Federal do Pará (UFPA)
Tucuruí – PA – Brasil

Abstract. Heavy rainfall threatens public safety and mining operations in Parauapebas, Pará (Eastern Amazon). We propose a compact pipeline that uses Bayesian optimization and XGBoost to predict days with rainfall exceeding 20 mm. The model seeks to maximize the F1 of the positive class under class imbalance, using daily data from CHIRPS and ERA5 (1981–2023). Temporal validation reflects real-world use and avoids data leaks. The classifier achieved an AUC-ROC of 0.927 and an AUC-PR of 0.488 in temporal testing (20%), showing potential for application in low-cost early warning systems.

Resumo. As chuvas intensas ameaçam a segurança pública e as operações de mineração em Parauapebas, Pará (Amazônia Oriental). Propomos um pipeline compacto que utiliza otimização Bayesiana e XGBoost para prever dias com precipitação superior a 20 mm. O modelo busca maximizar o F1 da classe positiva sob desequilíbrio de classes, utilizando dados diários do CHIRPS e ERA5 (1981–2023). A validação temporal reflete o uso real e evita vazamentos de dados. O classificador alcançou AUC-ROC de 0,927 e AUC-PR de 0,488 em teste temporal (20%), mostrando potencial de aplicação em sistemas de alerta precoce de baixo custo.

1. Introdução

Na Amazônia Oriental, o risco hidrometeorológico aumenta com a ocorrência de chuvas fortes, impactando diretamente a população das cidades e a logística de grandes projetos, como a mineração em Parauapebas, no estado do Pará. A habilidade de prever dias com chuvas intensas (acima de 20 mm) é fundamental para o planejamento de drenagem, proteção civil e gestão de ativos.

Os modelos de aprendizado de máquina (AM) têm progredido na identificação de extremos pluviométricos, sendo o XGBoost notável por sua robustez, interpretabilidade relativa (importância de atributos) e bom desempenho em situações tabulares. No entanto, sua utilidade operacional é condicionada por três desafios práticos: (i) prevenir vazamentos temporais na validação; (ii) lidar com o desbalanceamento de classes (eventos raros); e (iii) ajustar os hiperparâmetros de forma eficiente e alinhada à métrica de interesse.

Este estudo apresenta um classificador que prevê dias de chuva intensa em Parauapebas-PA, utilizando XGBoost, validação temporal consciente e otimização bayesiana. As contribuições mais relevantes são:

- Estrutura sucinta e reproduzível: com preditores de baixo custo (CHIRPS e ERA5), apropriada para a operação prática regional;
- Validação temporal explícita: prevenir vazamentos e representar o uso real ao longo da linha do tempo;
- Ajuste orientado à classe positiva: com foco em F1 e PR-AUC em meio a um forte desbalanceamento.

A avaliação em divisão temporal retida (20%) mostra um desempenho discriminativo consistente (ROC-AUC 0,927; PR-AUC 0,488), indicando a possibilidade de ser incorporada a sistemas de alerta precoce.

2. Trabalhos Relacionados

O XGBoost destaca-se em tarefas ambientais e hidrometeorológicas por sua robustez, interpretabilidade relativa e desempenho em dados tabulares [Chen and Guestrin 2016]. A otimização bayesiana, por sua vez, oferece uma busca eficiente de hiperparâmetros e reduz o risco de sobreajuste, sendo amplamente adotada em pipelines modernos de aprendizado de máquina [Snoek et al. 2012]. Na avaliação de classificadores em cenários de desbalanceamento, o uso combinado de F1, PR-AUC e ROC-AUC é recomendado para uma análise mais realista da capacidade discriminativa [Davis and Goadrich 2006].

Em relação às fontes de dados, o CHIRPS fornece precipitação diária com ampla cobertura global e boa consistência em regiões tropicais [Funk 2015], enquanto o ERA5 disponibiliza reanálises atmosféricas de alta resolução e qualidade reconhecida [Hersbach 2020]. Estudos recentes reforçam o potencial do aprendizado de máquina para previsão de eventos pluviométricos. [Wani 2024] evidenciaram o bom equilíbrio entre acurácia e custo computacional do XGBoost em relação a modelos profundos, enquanto [Oliveira 2023] demonstraram sua eficácia na previsão de chuvas intensas na Amazônia utilizando CHIRPS e ERA5, alinhando-se ao contexto deste estudo.

3. Dados e Métodos

Área de estudo e dados. Parauapebas-PA (06°04'S, 49°54'W), localizada na Amazônia Oriental, possui um padrão monossazonal de chuvas, com precipitações concentradas de dezembro a maio.¹ O intervalo de tempo analisado foi de 1981 a 2023. Séries diárias de precipitação CHIRPS v2.0 e reanálises ERA5 (ECMWF) foram empregadas para a temperatura média (2 m), umidade relativa e velocidade do vento (10 m), com dados interpolados para a coordenada central do município (SIRGAS 2000 / UTM 22S, EPSG:31982).

Tabela 1. Variáveis utilizadas.

Variável (Unid.)	Fonte
Precipitação (mm/dia)	CHIRPS v2.0
Temperatura (°C)	ERA5
Umidade relativa (%)	ERA5
Vento (m/s)	ERA5

Alvo. Chuva forte (>20 mm/dia) conforme INMET/CEMADEN → Classe 1; demais casos → Classe 0 ($\approx 8\%$ positivos).

Engenharia de atributos. Lags de precipitação ($t-1$, $t-7$), acumulados (3 e 7 dias) e defasagens de 1 dia para variáveis atmosféricas, totalizando 10 preditores, normalizados por *z-score*.

Divisão temporal. 15.708 amostras diários (1981–2023): 80 % identificação (1981–2016) e 20 % teste (2017–2023). Aplicou-se validação cruzada temporal (*Time-SeriesSplit*, $k=5$) para evitar vazamento.

¹Mapa público da área de estudo: <https://www.openstreetmap.org/#map=10/-6.07/-49.90>

Modelo e ajuste. XGBoost (*binary:logistic*) com ponderação $scale_pos_weight = \frac{n_{neg}}{n_{pos}}$. A otimização bayesiana (*gp_minimize*) maximizou o F1 da classe positiva sobre validação temporal. Espaço de busca: profundidade [3–10]; *learning rate* [0.01–0.3]; estimadores [100–600]; *subsample* e *colsample_bytree* [0.5–1]; *min_child_weight* [1–10]; regularizações L_1/L_2 [0–10].

Métricas. F1, precisão, revocação, PR-AUC e ROC-AUC no teste temporal. Enquanto a ROC-AUC avalia a separabilidade global, as métricas F1 e PR-AUC priorizam o desempenho da classe positiva em situações de desbalanceamento.

4. Resultados e Discussão

Desbalanceamento e desempenho. A série exibe uma predominância acentuada da classe 0 (cerca de 92%), um traço característico de eventos extremos. O desbalanceamento foi compensado pelo parâmetro *scale_pos_weight*, o que estabilizou o aprendizado. A distribuição de classes e a matriz de confusão normalizada obtidas no teste temporal (2017–2023) são exibidas na Figura 1.

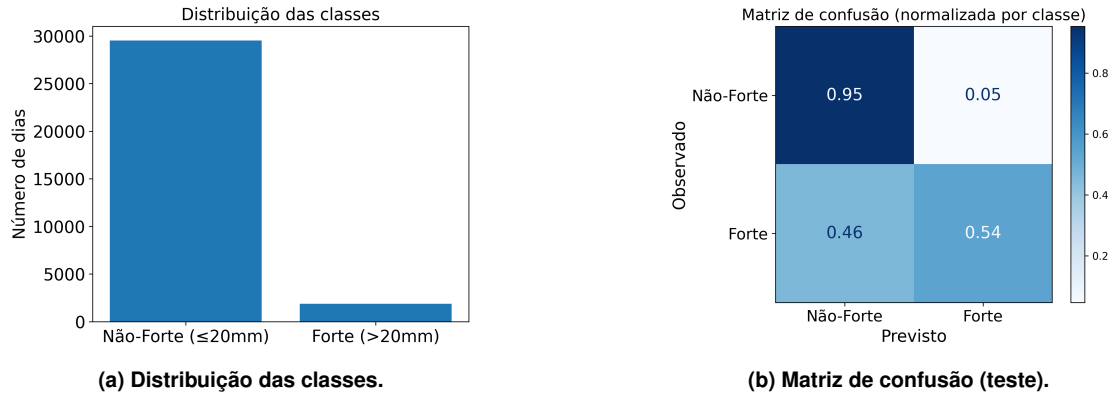


Figura 1. Desbalanceamento e desempenho por classe. Fonte: autoria própria.

Curvas ROC e PR. O modelo obteve um ROC-AUC de 0,927 e um PR-AUC de 0,488, valores considerados elevados em relação ao desbalanceamento. A Figura 2 sintetiza a habilidade discriminativa: a ROC exibe uma separação significativa entre as classes, enquanto a PR demonstra uma precisão consistente mesmo com uma baixa frequência da classe positiva..

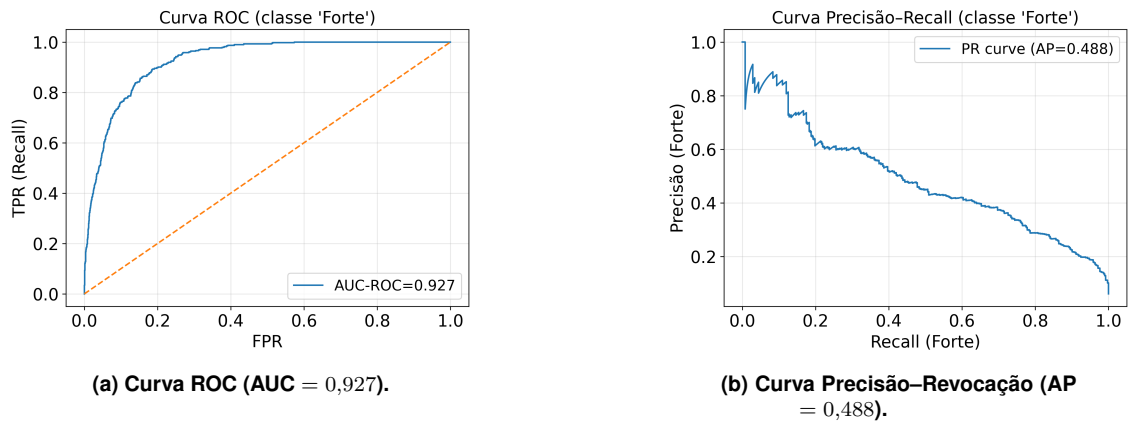


Figura 2. Curvas ROC e PR no conjunto de teste. Fonte: autoria própria.

Importância de variáveis. Umidade relativa, precipitação em $t-1$ e temperatura média destacaram-se como mais relevantes, indicando a influência de condições locais imediatas na deflagração de extremos, em concordância com [Oliveira 2023, Wani 2024].

Comparação com literatura. Os resultados superam o intervalo típico de ROC-AUC (0,80–0,88) relatado em estudos recentes de previsão de chuva com aprendizado de máquina [Wani 2024], demonstrando que a combinação de validação temporal e otimização bayesiana aprimora a estabilidade e a generalização [Oliveira 2023].

Aplicabilidade prática. O modelo é simples e de baixo custo computacional, podendo ser integrado a rotinas municipais ou minerárias de alerta hidrometeorológico para antecipar risco de alagamentos e interrupções operacionais.

5. Conclusão

Foi apresentado um pipeline eficiente baseado em XGBoost, utilizando otimização bayesiana e validação temporal, para prever com antecedência dias de chuvas intensas em Parauapebas, PA. Apesar do forte desbalanceamento, o modelo obteve um ROC-AUC de 0,93 e um PR-AUC de 0,49, mostrando uma boa capacidade de discriminação com preditores de baixo custo. A metodologia apresenta possibilidades de ser incorporada a sistemas de alerta hidrometeorológico locais. Como trabalhos futuros, sugere-se a utilização de técnicas de balanceamento (como SMOTE e class weights), a incorporação de índices climáticos de grande escala (como ENSO e MJO) e a comparação com redes LSTM e GRU.

Referências

- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 785–794.
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proc. 23rd Int. Conf. on Machine Learning*, pages 233–240.
- Funk, C. e. a. (2015). The climate hazards infrared precipitation with stations (chirps). *Scientific Data*, 2:150066.
- Hersbach, H. e. a. (2020). The era5 global reanalysis. *Q. J. R. Meteorol. Soc.*, 146:1999–2049.
- Oliveira, E. C. L. e. a. (2023). Precipitation forecasting: From geophysical aspects to machine learning applications. *Frontiers in Climate*, 5.
- Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25.
- Wani, O. A. e. a. (2024). Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the north-western himalayas. *Scientific Reports*, 14.