COMPARATIVO DE ALGORITMOS DE *MACHINE LEARNING* APLICADOS NA PREVISÃO DE ALAGAMENTOS

Melissa Sirqueira Pereira[1]; Selmo Eduardo Rodrigues Júnior[2]

Resumo

Este projeto pretende comparar 5 algoritmos computacionais aplicados na previsão de alagamentos, ancorado em modelos de aprendizado de máquina, a saber o *Random Forest, XGboost*, *Linear Discriminant Analysis* (LDA), Decision Tree e *K-Nearest Neighbors* (KNN) na cidade de Imperatriz, no estado do Maranhão. Com a intenção de promover estudos regionais, propõe-se a utilização de abordagens de aprendizado de máquina, alimentando estes algoritmos com dados de alagamentos, informações meteorológicas associadas ou informações físicas sobre o rio Tocantins, recolhidas localmente, através da extração de notícias disponíveis na internet e dispostas pela prefeitura municipal. Para tanto, a coleta e tratamento de dados de ocorrências de alagamentos foi datado entre 2020 e 2024. Buscando reduzir a distorção nos resultados, foi aplicado um tratamento de dados para padronizar a base de dados. Com o acervo de dados organizado, identificou-se as *features* de maior importância e utilizou-se *feature engineering* para otimizar a entrada dos modelos. Por fim, os algoritmos foram submetidos à validação cruzada e avaliados pelas métricas de *accuracy, precision, recall* e F1 score. Modelos de melhor desempenho individual foram subsequentemente combinados por meio de *Ensemble Learning* e reavaliados sob as mesmas métricas. O modelo *Random Forest* demonstrou performance superior em todas as métricas avaliadas. Os resultados foram disponibilizados em uma interface *localhost* para demonstração.

Palavras-chave: Previsão; Machine Learning; Alagamentos; Imperatriz.
Financiamento: Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão - FAPEMA

[1] Estudante do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail:melissa.sirqueira@acad.ifma.edu.br.

[2] Professor Dr. do Curso de Engenharia Elétrica do IFMA do Campus Imperatriz; E-mail: selmo.junior@ifma.edu.br.

Introdução

Ao longo da história, inúmeras inundações ocorreram por diferentes motivos, o que impulsionou a realização de estudos para compreender esse fenômeno em detalhe (O'Connor & Costa, 2004). Modelos matemáticos têm se mostrado uma alternativa promissora, pois utilizam dados históricos para medir ou prever enchentes, sendo amplamente aplicados em várias regiões (Valipour; Banihabib & Behbahani, 2013).

As catástrofes ambientais se caracterizam pela imprevisibilidade e pelos danos sociais, humanos e econômicos que provocam. A falta de controle ou de medidas preventivas pode gerar atrasos e custos elevados em reparos que poderiam ser evitados. Um exemplo é o aumento da frequência de inundações em rios e áreas costeiras, fenômeno associado ao aquecimento global e ao desequilíbrio climático. Essas condições se agravaram a partir da década de 1950 e continuam em escala crescente (IPCC, 2023).

Nesse contexto, é fundamental adotar ações que tornem as cidades mais preparadas para lidar com eventos extremos, protegendo vidas e reduzindo impactos estruturais (Zieba *et al.*, 2020). Estudos recentes mostram que a modelagem matemática desempenha um papel central nesse processo, e técnicas de *machine learning* (ML) vêm sendo aplicadas para prever enchentes e minimizar danos (Ghorpade et al., 2021).

Segundo Mosavi, Ozturk & Chau (2018), os algoritmos de ML mais utilizados na previsão de inundações incluem Redes Neurais Artificiais (RNAs), modelos neuro-*fuzzy,* ANFIS (*Adaptive Neuro-Fuzzy Inference System*), Máquinas de Vetores de Suporte (SVM), Redes Neurais *Wavelet* (WNN) e *Perceptrons* Multicamadas (MLP). A partir dessas abordagens, foram desenvolvidos modelos derivados que buscam maior precisão de acordo com o contexto e com a disponibilidade de dados (Seydi *et al.*, 2023).

Apesar dos avanços, ainda existem limitações importantes: os modelos podem ser afetados por parâmetros pouco conhecidos e pela falta de informações quantitativas confiáveis. Isso mostra a necessidade de novas alternativas que preencham essas lacunas e tornem a previsão de enchentes mais eficaz (Cerna; Guyeux & Laiymani, 2022).

Linardos *et al*. (2022) observam que pesquisas com *machine learning* e *deep learning* já fornecem alternativas em diversas fases da gestão de desastres, desde a preparação até a recuperação pós-evento. Isso reforça a ideia de que tais tecnologias podem apoiar todas as etapas do gerenciamento de riscos.

Por fim, a seleção criteriosa de modelos é essencial para identificar a abordagem mais eficaz em cada contexto. Um exemplo prático é o estudo de Facco *et al.* (2020), que avaliou

diferentes modelos de ML para prever enchentes em Santa Maria, levando em consideração não apenas os dados disponíveis, mas também as particularidades da região e suas necessidades de gestão de riscos.

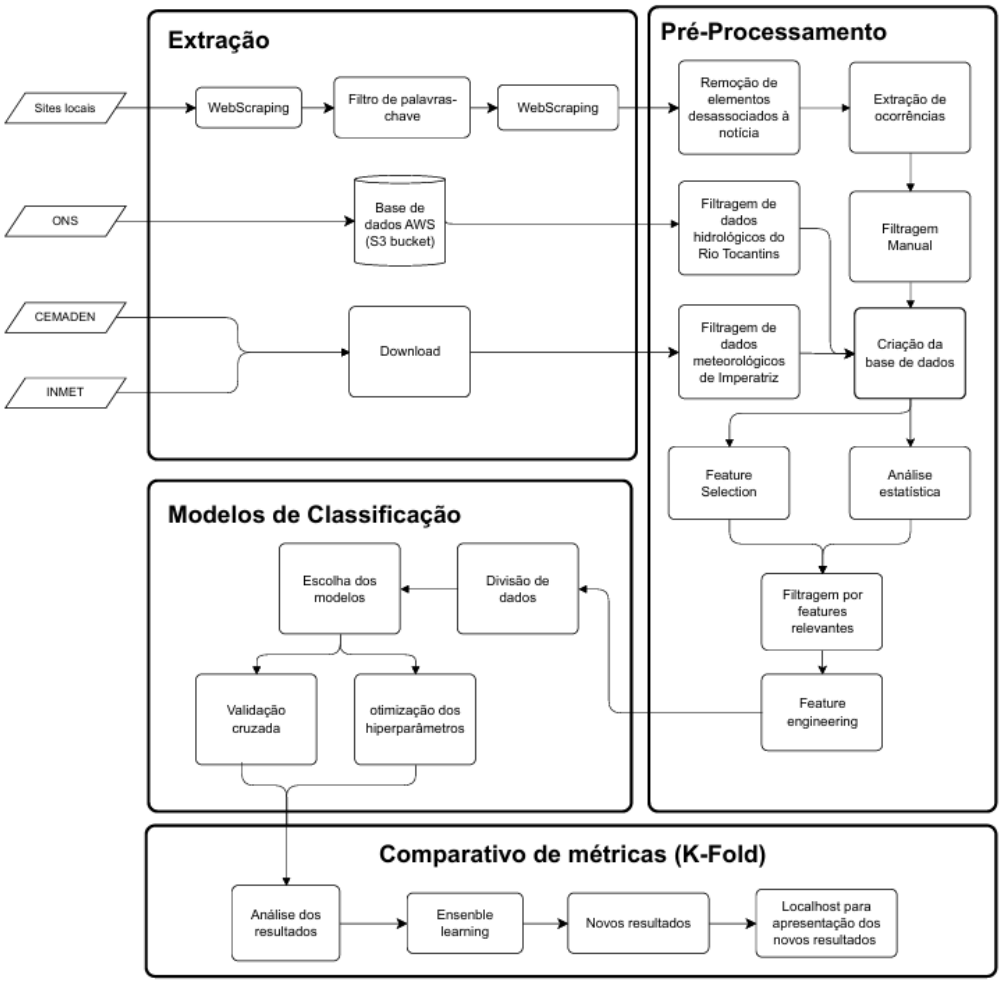
Metodologia

Empregou-se o uso de *softwares* que utilizam a aplicação *Jupyter Notebook*, entre elas o *Google Colab* que possui a própria unidade de processamento, poupando custo computacional. Ademais, utilizou-se a linguagem *Python* devido às diversas bibliotecas voltadas para extração de dados, *Machine Learning* e Inteligência Artificial.

A base de dados sobre eventos meteorológicos foi composta por informações exibidas em diferentes sites governamentais e a confirmação de alagamentos foi montada a partir de sites de notícias locais e da prefeitura do município de Imperatriz. Foram consultadas múltiplas fontes oficiais: INMET, CEMADEN e ONS, além da Defesa Civil de Imperatriz-MA, que forneceu relatórios sobre inundações. Parte dos dados foi obtida de forma não automatizada, e os do ONS estavam disponíveis em ambiente AWS (S3 *bucket*).

A Figura 1 apresenta o fluxograma com os passos seguidos ao longo da pesquisa. O processo inicia-se com a extração das bases de dados, que corresponde à primeira etapa do estudo. Em seguida, os dados passam pelo pré-processamento, etapa essencial para eliminar informações que poderiam distorcer os resultados da fase seguinte: a seleção de features, para entender a importância na ocorrência de alagamento. A fase seguinte se trata da escolha dos modelos iniciais e o teste inicial, na sequência está a fase final, na qual os modelos passam pelo *Ensemble Learning*, são testados novamente para apresentar os resultados finais.

Figura 1: Fluxograma da Metodologia de Pesquisa



Fonte: Autoria própria, 2025

Diante da escassez de registros oficiais, adotou-se a extração de notícias locais via *Web Scraping*, dividida em duas etapas: (i) coleta de títulos com filtragem via Processamento de Linguagem Natural (NLP) usando palavras-chave como “alagamento” e “chuva”; (ii) coleta detalhada das descrições. Em seguida, realizou-se uma classificação manual, atribuindo valor 1 para casos confirmados de alagamentos naturais em Imperatriz-MA e 0 para não ocorrências.

Para padronizar as informações, foram removidas variáveis com mais de 50% de valores ausentes, além de duplicatas. As séries foram convertidas para periodicidade diária, agregando médias, e unificadas em uma única coluna de data. Esse tratamento garantiu a remoção de ruídos e tornou os dados consistentes para a análise.

A base inicial continha muitas variáveis, algumas pouco relevantes ou externas à área de estudo. Para reduzir a dimensionalidade, aplicaram-se técnicas de seleção de variáveis:

- Correlação de Spearman: adequada para variáveis contínuas e ordinais, permitindo identificar fatores diretamente associados à ocorrência de alagamentos (Siegel, 1975).
- *Random Forest*: além de classificador, também foi utilizado para medir a importância das variáveis, identificando as features mais relevantes para o problema (Breiman, 2001; Conceição, 2022; Salman, 2024). A análise foi feita em 70% dos dados, sem ajustes de parâmetros, com o objetivo exclusivo de seleção de variáveis.

Foram criadas novas variáveis derivadas das existentes. Um exemplo foi o cálculo da chuva acumulada nos últimos 7 dias, fornecendo um indicativo mais robusto de períodos críticos e melhorando a capacidade de generalização dos modelos.

Para o treinamento dos modelos, cinco algoritmos foram selecionados por suas características complementares:

- *Linear Discriminant Analysis* (LDA) - modelo estatístico linear clássico.
- *Decision Tree* – interpretável e útil para extrair regras.
- *Random Forest* – robusto ao combinar múltiplas árvores.
- *XGBoost* – baseado em *Gradient Boosting*, reconhecido por alto desempenho preditivo.
- *K-Nearest Neighbors* (KNN) - não paramétrico, baseado em proximidade entre amostras.

Após o treinamento inicial, realizou-se otimização de hiperparâmetros por meio de *Randomized Search Cross-Validation*, explorando combinações aleatórias de parâmetros como profundidade de árvore, número de estimadores, taxa de aprendizado e vizinhos do KNN, exibidos na Tabela 1.

Tabela 1 - Definição de hiperparâmetros

Modelos					
Modelos	**LDA**	**Random Forest**	**Decision Tree**	**XGBoost**	**KNN**
N_estimators	-	100 a 1000	-	100 a 1000	-
Max_depth	-	5 a 100	5 a 100	3 a 50	-
Min_samples_split	-	2 a 20	2 a 20	-	-
Min_samples_leaf	-	1 a 20	1 a 20	-	-
Learning_rate	-	-	-	0.001 a 0.5	-
Subsample	-	-	-	0.3 a 0.7	-
Colsample_bytree	-	-	-	0.3 a 0.7	-
N_neighbors	-	-	-	-	1 a 50
Weights	-	-	-	-	“uniform”, “distance”
P	-	-	-	-	1, 2

Fonte: Autoral, 2025

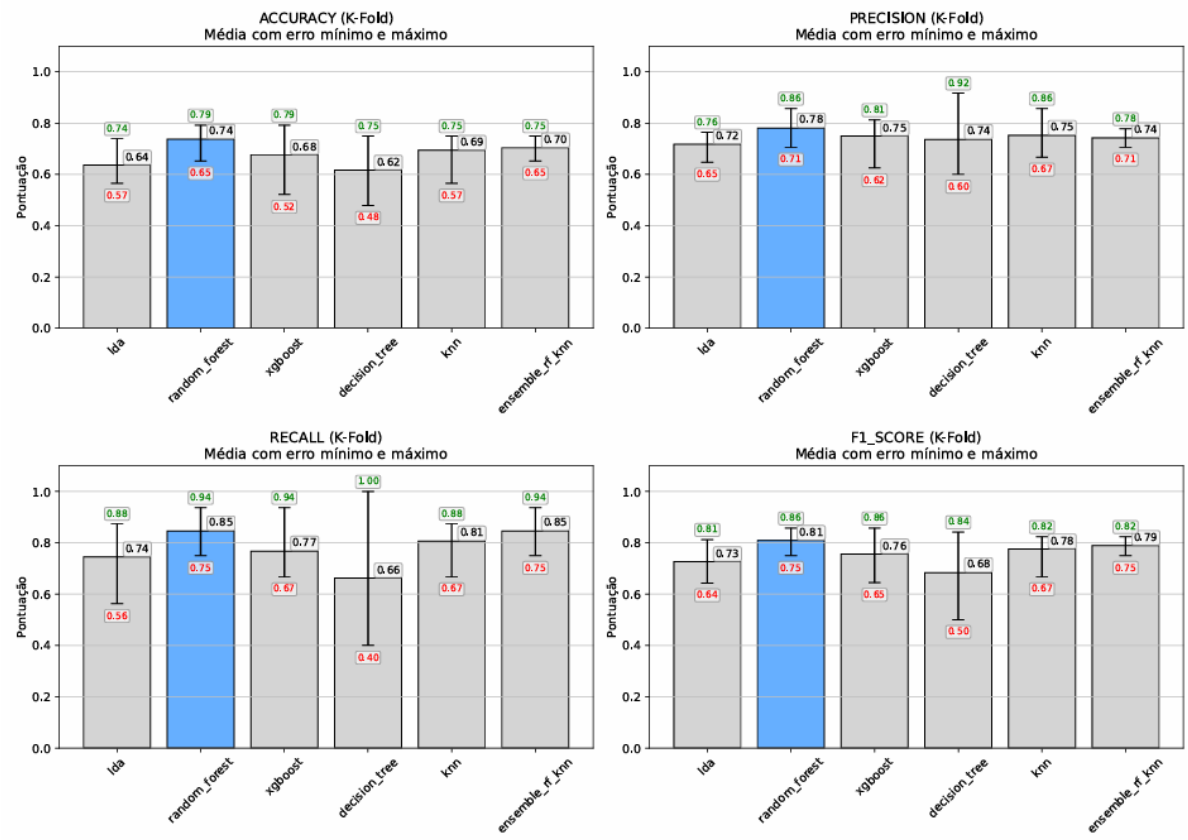
Para avaliar a robustez, utilizou-se validação cruzada *K-Fold* (k=5), garantindo múltiplas rodadas de treino e teste. As métricas de desempenho incluíram *acurácia*, precisão, *recall* e *F1-score*, cada uma analisada em média, erro mínimo e máximo.

Por fim, foi aplicada a técnica de *Ensemble Learning*, que combina diferentes modelos para reduzir viés e variância, aumentando a generalização (Jadama, 2024). Após análise comparativa, os modelos *Random Forest* e KNN apresentaram melhor desempenho. Foi implementado o método *Bagging* (*Bootstrap Aggregating*), no qual os modelos são treinados em subconjuntos gerados por amostragem com reposição, e suas previsões são combinadas via *hard voting*. Esse procedimento buscou ganhos em *acurácia* e *F1-score*, fundamentais diante do número limitado de ocorrências positivas de alagamentos.

Resultados e Discussão

Após as etapas de coleta e pré-processamento dos dados foi mensurada a importância das *features* que o modelo necessita, e em seguida as técnicas apresentadas no tópico "modelos de classificação" da Figura 1, onde os resultados dessas operações estão representados na Figura 2, que indica que o modelo com maiores pontuações médias dentre as métricas de avaliação, *Random Forest*, está destacado em azul, seguido pelo *Ensemble Learning* entre o *Random Forest* e o modelo KNN.

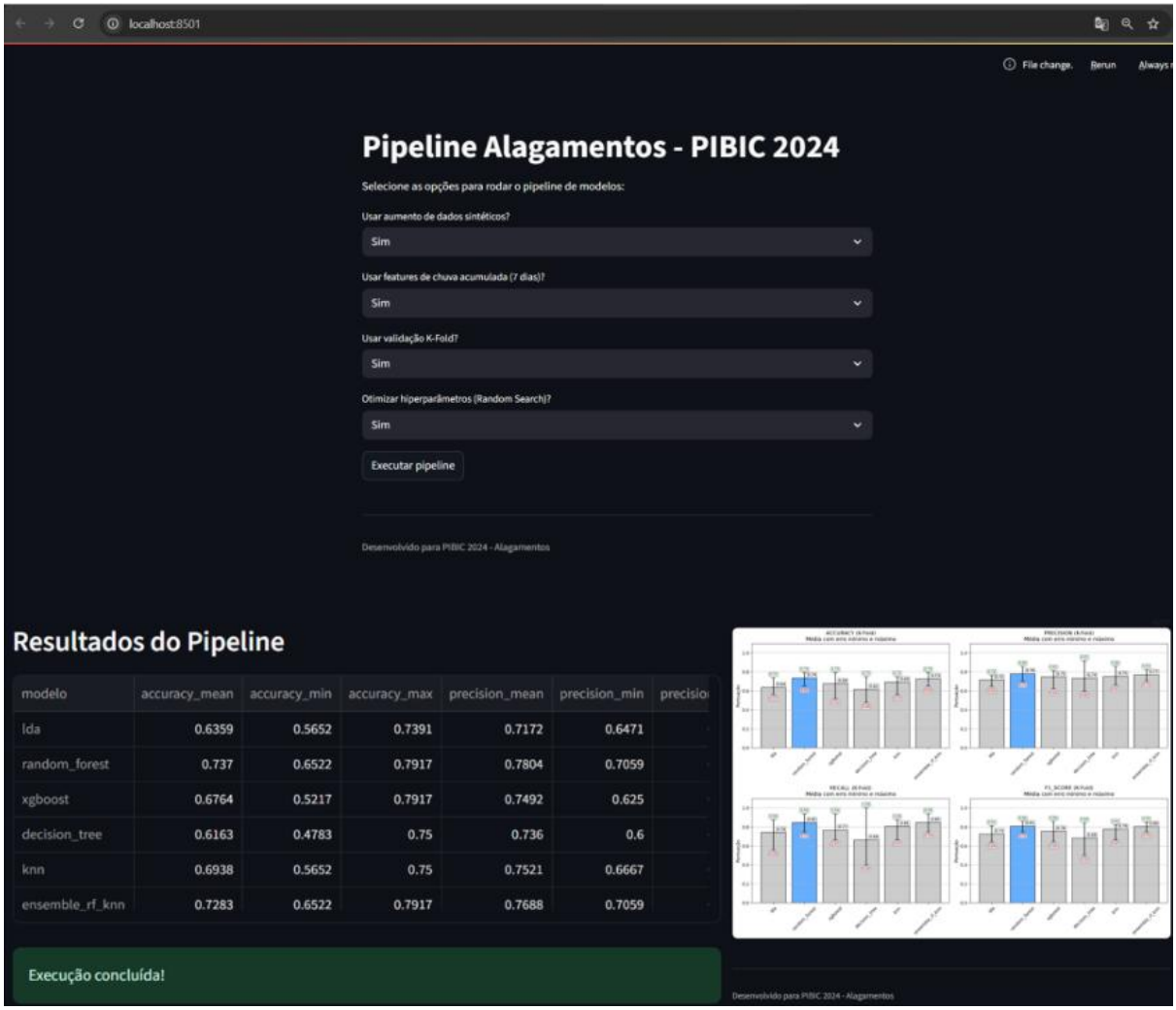
Figura 2: Análise das métricas de avaliação



Fonte: Autoria própria, 2025

Na Figura 3, apresenta-se o layout do servidor local, exibindo a tabela e o gráfico como resultado obtidos Figura 2.

Figura 3: Layout do servidor local, com tabela e gráfico dos resultados da Figura 2.



Fonte: Autoria própria, 2025

Conclusão

A análise desenvolvida identifica as variáveis meteorológicas e hidrológicas mais relevantes para a ocorrência de alagamentos em Imperatriz-MA e as aplica em diversos modelos de aprendizado de máquina. Após *feature engineering*, ajuste de hiperparâmetros e aplicação de *Ensemble Learning*, observa-se que algoritmos mais complexos, principalmente o *Random Forest*, apresentam melhor desempenho e menor variação entre os *folds* da validação cruzada. A combinação *Random Forest* + *KNN* também demonstra robustez, enquanto o *XGBoost* mostra alta precisão, porém com maior oscilação, indicando necessidade de ajustes finos. Modelos mais simples, como *Decision Tree, KNN* e *LDA*, não se mostram competitivos devido à limitada capacidade de generalização. A escassez de dados e a ausência de algumas

medições climáticas limitam o treinamento, refletindo em variações nos resultados. A natureza complexa e não linear dos alagamentos, envolvendo múltiplos fatores ambientais e antrópicos, dificulta previsões precisas.

Apesar dessas limitações, a metodologia desenvolvida contribui para o avanço do uso de aprendizado de máquina na gestão de riscos de eventos extremos. Para trabalhos futuros, recomenda-se ampliar a base de dados, aplicar técnicas de balanceamento de classes, integrar variáveis espaciais e combinar modelos físicos de hidrologia com modelos de aprendizado de máquina híbridos, aumentando a precisão e a generalização das previsões.

Agradecimentos

Agradeço à FAPEMA pelo apoio financeiro que tornou este estudo possível, ao INMET, CEMADEN, ONS e à Defesa Civil de Imperatriz pela disponibilização de dados, ao pesquisador Pablo Francisco Melo Bezerra e a todos que contribuíram para a coleta, organização e análise das informações, viabilizando o desenvolvimento desta pesquisa.

Referências

BREIMAN, L. **Random forests. Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CERNA, S.; GUYEUX, C.; LAIYMANI, D. **The usefulness of NLP techniques for predicting peaks in firefighter interventions due to rare events.** ***Neural Computing and Applications***, v. 34, p. 10117–10132, 2022. DOI: https://doi.org/10.1007/s00521-022-06996-x.

CONCEIÇÃO, L. S. M. **A utilização do random forest para prever a inflação.** 2022. Monografia (Graduação em Economia) – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2022.

FACCO, M.; CAMPOS, M. A. de A. de; VARGAS, D. dos S.; SILVEIRA, R. B.; BISOGNIN, C. **Algoritmos de Machine Learning Aplicados na Ocorrência de Chuvas na Cidade de Santa Maria.** *Ciência e Natura*, v. 42, p. e28, 2020. DOI: https://doi.org/10.5902/2179460X40537.

GHORPADE, P. et al. **Flood Forecasting Using Machine Learning: A Review.** In: *INTERNATIONAL CONFERENCE ON SMART COMPUTING AND COMMUNICATIONS (ICSCC)*, 8., 2021, [s.l.]. *Anais* [...]. [s.l.]: IEEE, 2021. p. 32-36. DOI: https://doi.org/10.1109/ICSCC51209.2021.9528099.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (IPCC). **Weather and Climate Extreme Events in a Changing Climate.** In: *CLIMATE CHANGE 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 2023. p. 1513-1766. DOI: https://doi.org/10.1017/9781009157896.013.

LINARDOS, V.; DRAKAKI, M.; TZIONAS, P.; KARNAVAS, Y. L. **Machine Learning in Disaster Management: Recent Developments in Methods and Applications.** *Machine Learning and Knowledge Extraction*, v. 4, n. 2, p. 446-473, 2022. DOI: https://doi.org/10.3390/make4020020.

MOSAVI, A.; OZTURK, P.; CHAU, K. W. **Flood Prediction Using Machine Learning Models: Literature Review.** *Água*, v. 10, p. 1536, 2018. DOI: https://doi.org/10.3390/w10111536.

O'CONNOR, J. E.; COSTA, J. E. **The world's largest floods, past and present—Their causes and magnitudes**. *U.S. Geological Survey Circular*, n. 1254, 2004. 13 p.

SALMAN, H.; KALEKACH, A.; STEITI, A. **Random Forest Algorithm Overview.** *Babylonian Journal of Machine Learning*, p. 69-79, 2024. DOI: https://doi.org/10.58496/BJML/2024/007.

SEYDI, S. T. et al. **Comparison of Machine Learning Algorithms for Flood Susceptibility Mapping.** *Remote Sensing*, v. 15, n. 1, p. 192, 2023. DOI: https://doi.org/10.3390/rs15010192.

SIEGEL, S. ***Estatística não-paramétrica: para as ciências do comportamento*.** São Paulo: McGraw-Hill do Brasil, 1975.

VALIPOUR, M.; BANIHABIB, M. E.; BEHBAHANI, S. M. R. **Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir.** *Journal of Hydrology*, v. 476, p. 433-441, 2013. DOI: https://doi.org/10.1016/j.jhydrol.2012.11.017.

ZIĘBA, Z. et al. **Built Environment Challenges Due to Climate Change.** *IOP Conference Series: Earth and Environmental Science*, v. 609, p. 012061, 2020. 6th World Multidisciplinary Earth Sciences Symposium, 7-11 Sept. 2020, Prague, Czech Republic. DOI: https://doi.org/10.1088/1755-1315/609/1/012061