XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

APPLYING MACHINE LEARNING TECHNIQUES FOR BINARY LEAK DETECTION IN WATER DISTRIBUTION SYSTEMS: A COMPARATIVE

Carreño-Alvarado E. P.[1]*; Reynoso-Meza G.*[2]*; Scapulatempo Fernandes C. V.*[3] & Rossa R. R.[4].

Abstract: Sanitation, hygiene, and access to clean water are fundamental human rights. However, water distribution networks deteriorate over time due to aging infrastructure, failures, and increasing water demand driven by population growth. This degradation leads to water losses and operational inefficiencies, making efficient resource management essential and water system management increasingly complex.

This study compares three Artificial Intelligence (AI) Machine Learning (ML) techniques - K-Nearest Neighbors (KNN), Random Forest, and XGBoost - for binary leak detection in water distribution systems. A hydraulic model based on the Hanoi network was used to simulate both normal and leak conditions. Seven scenarios (four without leaks and three with leaks) were considered, generating twelve training/testing combinations using a 70/30% controlled split.
The results demonstrate that AI models can classify operational data to detect leaks, with ensemble methods (Random Forest and XGBoost) generally outperforming KNN in terms of accuracy and consistency. This research highlights the potential of AI-based tools to help water utilities improve operational efficiency, reduce non-revenue water, and enhance the sustainability of distribution networks. By integrating machine learning with hydraulic simulation, the study contributes to practical solutions for one of the most pressing challenges in water management. The findings suggest that AI models can enhance leak detection capabilities, providing a valuable complement to traditional monitoring methods and enabling proactive water loss management strategies.

Resumo: Saneamento, higiene e acesso à água potável são direitos humanos fundamentais. No entanto, as redes de distribuição de água se deterioram ao longo do tempo devido ao envelhecimento da infraestrutura, falhas e aumento da demanda hídrica impulsionado pelo crescimento populacional. Essa degradação leva a perdas de água e ineficiências operacionais, tornando essencial o gerenciamento eficiente dos recursos e a gestão dos sistemas hídricos cada vez mais complexa.

Este estudo compara três técnicas de Aprendizado de Máquina (Machine Learning – ML) baseadas em Inteligência Artificial (IA) — K-Nearest Neighbors (KNN), Random Forest e XGBoost — para a detecção binária de vazamentos em sistemas de distribuição de água. Um modelo hidráulico baseado na rede de Hanói foi utilizado para simular condições normais e com vazamentos. Foram

1) Posdoutoranda no Programa de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental – PPGERHA - UFPR. Av. Cel. Francisco H. dos Santos, 210 – Jardim das Américas – Bloco V – DHS – LME. Curitiba/PR. E-mail: elizabeth.pauline@ufpr.br
2)Professor pesquisador, Programa Programa de Pós-graduação em Engenharia de Produção e Sistemas (PPGEPS), Pontifícia Universidade Católica do Paraná (PUCPR), Rua Imaculada Conceição, 1155, Prado Velho, Curitiba, Paraná, Brazil. Faculty of engineering, Shinhu University, Nagano, Prefctura de Nagano. E-mail: greynosom@pucpr.br
[3] Professor, Departamento de Hidráulica e Saneamento – UFPR. Av. Cel. Francisco H. dos Santos, 210 – Jardim das Américas – Bloco V – DHS – LME. Curitiba/PR. E-mail: cris.dhs@ufpr.br
[4] Programa de Pós-Graduação em Engenharia de Recursos Hídricos e Ambiental – PPGERHA - UFPR. Av. Cel. Francisco H. dos Santos, 210 – Jardim das Américas – Bloco V – DHS – LME. Curitiba/PR. E-mail: rafaelrarossa@gmail.com

considerados sete cenários (quatro sem vazamentos e três com vazamentos), gerando doze combinações de treino/teste usando uma divisão controlada de 70/30%.

Os resultados demonstram que modelos de IA podem classificar dados operacionais para detectar vazamentos, com os métodos de ensemble (Random Forest e XGBoost) geralmente superando o KNN em termos de acurácia e consistência. Esta pesquisa destaca o potencial de ferramentas baseadas em IA para ajudar as operadoras de água a melhorar a eficiência operacional, reduzir perdas não contabilizadas e aumentar a sustentabilidade das redes. Ao integrar aprendizado de máquina com simulação hidráulica, o estudo contribui com soluções práticas para um dos desafios mais urgentes da gestão hídrica.

Palavras-Chave – Hydraulic Modeling, Water Loss Management, Supervised Learning Models.

INTRODUCTION

Water distribution systems (WDS) are crucial for providing clean water. However, water losses due to leaks represent a significant global issue, contributing to resource waste, economic losses, and environmental impacts. For example, in Brazil, data indicate the loss rate in water distribution (INO049) was 37,8% (SNIS-AE 2022), slightly better than the total revenue loss rate of 37,06% recorder in 2018 or the 39.21% recorded in 2017. In contrast, the distribution loss rate reached 38.45%, slightly worse than the 38.29% observed in 2017 (SNIS, 2018).

Recent advances in Artificial Intelligence (AI) offer promising solutions for improving the monitoring and management of WDS, particularly in leak detection, where AI has demonstrated effectiveness in various applications. This study aims to evaluate and compare the performance of three AI-based machine learning classification techniques: K-Nearest Neighbors (KNN), XGBoost, and Random Forest. for binary leak detection (i.e., determining whether a leak is present or not) using hydraulic simulation data. These methods were selected due to their proven effectiveness in classification tasks: KNN for its simplicity and interpretability (Cover & Hart, 1967), Random Forest for its robustness to overfitting (Breiman, 2001), and XGBoost for its high accuracy in handling structured data (Chen & Guestrin, 2016).

METHODOLOGY

The goal of this study is to detect leaks in water distribution networks early and accurately to reduce unaccounted-for water losses, improve operational efficiency, and minimize both environmental and economic impacts. The CRISP-DM methodology (Cross Industry Standard Process for Data Mining) will be used. It is a very standard methodology in data science and machine learning (Schröer et al., 2021), show in the Figure 1, was followed and adapted into four stages, as outlined below:

Stage I: Business Understanding, Data Understanding, and Data Preparation.

This initial stage involved the generation and preprocessing of hydraulic data. A benchmark dataset was used; all data have been compiled, checked, and labeled. The benchmark data were sourced from the LeakDB archive developed for the 1st International WDSA/CCWI 2018 Joint Conference.

Stage II: Modeling and Evaluation.

Machine learning models were trained and evaluated during this stage. Data preparation was iteratively refined as needed to improve model performance. Adjustments were made to ensure the best possible input quality.

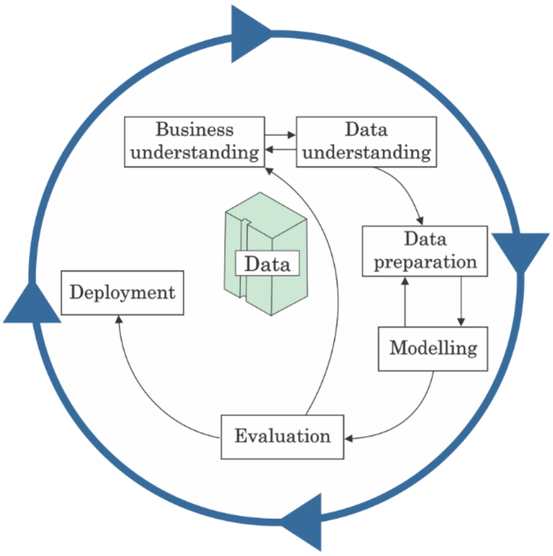
Stage III: Re-evaluation.

This stage focused on a critical analysis of results to determine whether any previous steps needed to be repeated or adjusted.

Stage IV: Deployment preparation.

This stage involved waiting for real-world data to finalize potential deployment.

Figure 1 – CRISP-DM Methodology



Three classical machine learning models from AI were selected: Random Forest, XGBoost, and K-Nearest Neighbors (KNN). These were chosen for their simplicity, interpretability, and performance. The model advantages are shown in Table 1. A summary of their main characteristics is provided below:

- **K-Nearest Neighbors (KNN):** A simple, instance-based supervised learning algorithm used for classification and regression. It predicts a new observation's label based on the majority vote or average of its “k” closest neighbors in the feature space (Cover & Hart, 1967).
- **Random Forest:** An ensemble learning method that builds multiple decision trees and combines their predictions through majority voting (for classification) or averaging (for regression). It introduces randomness via bootstrapped sampling and feature selection to reduce overfitting (Breiman, 2001).

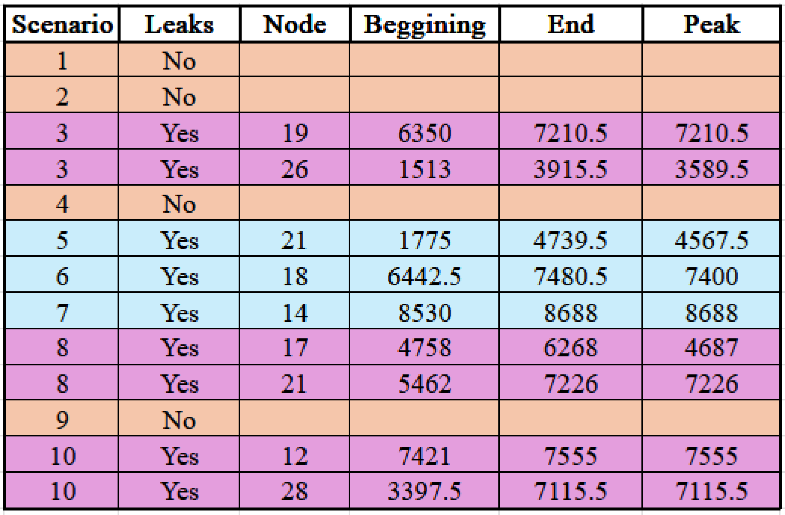
- **XGBoost:** An advanced gradient boosting framework that builds trees sequentially, where each tree corrects the errors of the previous ones. It optimizes a loss function using gradient descent, with strong performance and built-in regularization (Chen & Guestrin, 2016).

Table 1 – Advantages of the machine learning methods

KNN	Random Forest	XGBoost
Simple implementation with no explicit training phase. Dynamically adapts to new data. Effective for small, low-dimensional datasets.	Reduces overfitting via ensemble averaging and feature randomness. Handles high-dimensional data effectively. Provides feature importance metrics. Robust to outliers and noise.	State-of-the-art predictive performance. Built-in (L1/L2) regularization. Efficient handling of missing data. Scalable with parallel processing.

Hydraulic data included 10 different scenarios with characteristics detailed in Table 2. The simulations were implemented in Python using Anaconda, leveraging standard libraries, WNTR, the EPyT tool, and EPANET. Each scenario simulated a one-year period with 30-minute time steps, generating data on pressure, demand, and flow. These parameters were extracted from each scenario to build the dataset, with labels added to prepare it for use in training and testing the different models.

Table 2 – The scenarios from the benchmark



Scenario	Leaks	Node	Beggining	End	Peak
1	No				
2	No				
3	Yes	19	6350	7210.5	7210.5
3	Yes	26	1513	3915.5	3589.5
4	No				
5	Yes	21	1775	4739.5	4567.5
6	Yes	18	6442.5	7480.5	7400
7	Yes	14	8530	8688	8688
8	Yes	17	4758	6268	4687
8	Yes	21	5462	7226	7226
9	No				
10	Yes	12	7421	7555	7555
10	Yes	28	3397.5	7115.5	7115.5

To simplify the classification task, only scenarios with one leak and without leaks were considered. A total of seven scenarios were selected: four without leaks and three with one leak. This resulted in twelve unique training/testing combinations (with a 70/30% split), as shown in Table 3.

Table 3 – The scenarios from the benchmark

	Scenarios considered						
Combinations	Train					Test	
Combi_1	2	4	6	7	9	1	5
Combi_2	2	4	5	7	9	1	6
Combi_3	2	4	5	6	9	1	7
Combi_4	1	4	6	7	9	2	5
Combi_5	1	4	5	7	9	2	6
Combi_6	1	4	5	6	9	2	7
Combi_7	1	2	6	7	9	4	5
Combi_8	1	2	5	7	9	4	6
Combi_9	1	2	5	6	9	4	7
Combi_10	1	2	4	6	7	9	5
Combi_11	1	2	4	5	7	9	6
Combi_12	1	2	4	5	6	9	7

🟩 No Leak

🟥 Leak

Hydraulic data were compiled into datasets with labeled instances, each including the timestamp and a leak label (1 = leak, 0 = no leak). Each combined data from multiple nodes over time, capturing dynamic system behavior.

Twelve combinations were generated to evaluate model generalizability, balancing classes in both training and testing sets (70/30%).

All models were implemented in Python using Anaconda, leveraging standard libraries. The computations were executed on a computer with the following specifications:

- Operating System: Windows 11
- Processor: AMD Ryzen 5 5500U with Radeon Graphics, 2.10 GHz, 6 cores, 12 logical processors
- RAM: 16 GB

Training times varied significantly:

- KNN: around 7 minutes
- XGBoost: around 21 minutes
- Random Forest: around 90 minutes

A noteworthy difference in computational cost between the models.

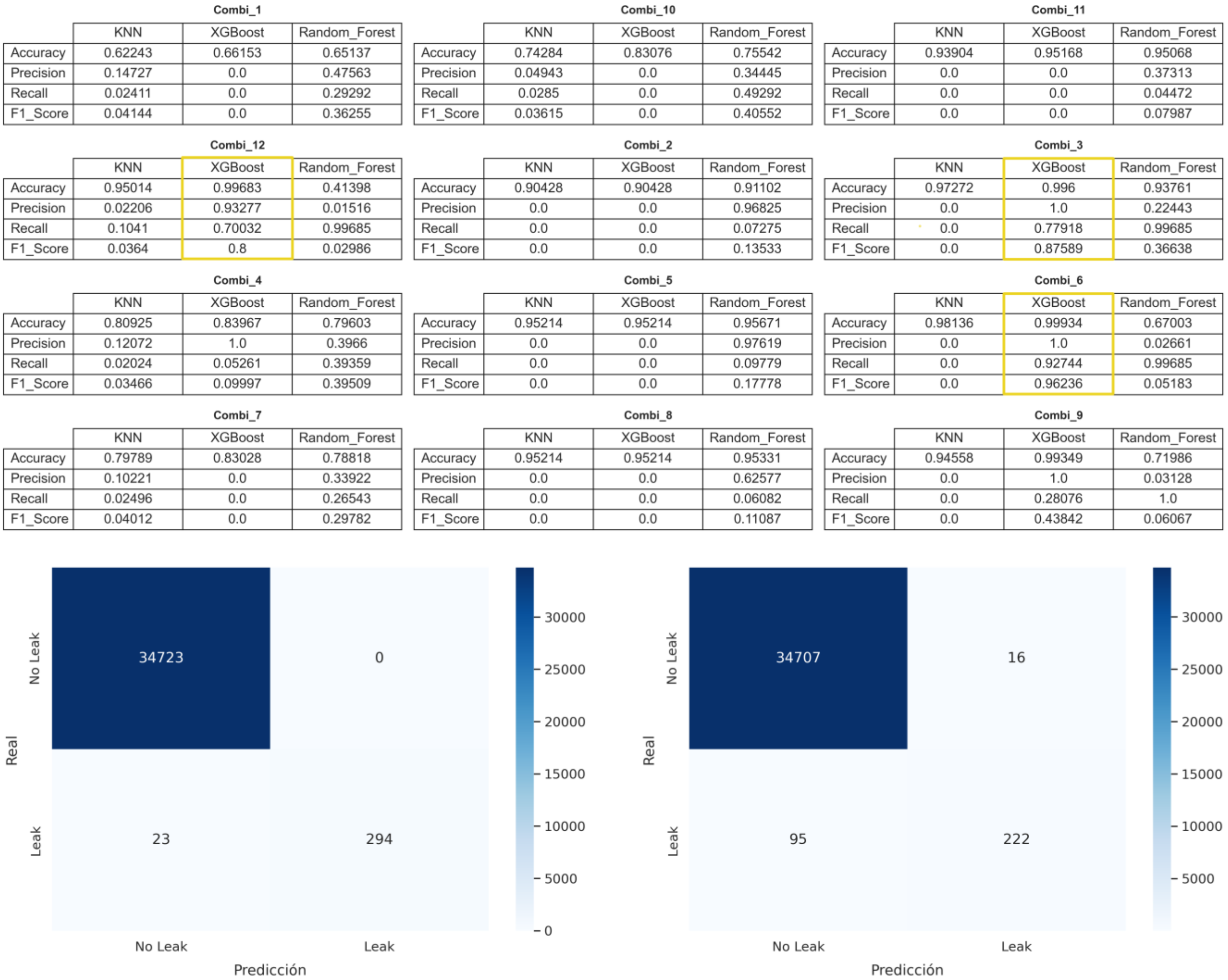
RESULTS

The models were evaluated using the following standard classification metrics: accuracy, precision, recall, F1-score and the respective confusion matrices.

Recall (also called Sensitivity or True Positive Rate) measures how well a model identifies all relevant positive cases. It calculates the proportion of actual positives correctly predicted as positive (finding all leaks in a water network). High Recall means fewer missed positives (Powers 2011). Precision (also called Confidence or True Positive Accuracy) evaluates how reliable the model's positive predictions are. It measures the proportion of predicted positives that are truly positive (leaks flagged by the model that are real leaks, not false alarms) (Powers 2011).

The F1-score (or F-measure) is the harmonic mean of Precision and Recall, emphasizing balance. A confusion Matrix is a tabular representation of true vs. predicted labels, detailing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Figure 2– Metrics of the different combinations and confusion matrices, of the different models.



Combi_1

	KNN	XGBoost	Random_Forest
Accuracy	0.62243	0.66153	0.65137
Precision	0.14727	0.0	0.47563
Recall	0.02411	0.0	0.29292
F1_Score	0.04144	0.0	0.36255

Combi_10

	KNN	XGBoost	Random_Forest
Accuracy	0.74284	0.83076	0.75542
Precision	0.04943	0.0	0.34445
Recall	0.0285	0.0	0.49292
F1_Score	0.03615	0.0	0.40552

Combi_11

	KNN	XGBoost	Random_Forest
Accuracy	0.93904	0.95168	0.95068
Precision	0.0	0.0	0.37313
Recall	0.0	0.0	0.04472
F1_Score	0.0	0.0	0.07987

Combi_12

	KNN	XGBoost	Random_Forest
Accuracy	0.95014	0.99683	0.41398
Precision	0.02206	0.93277	0.01516
Recall	0.1041	0.70032	0.99685
F1_Score	0.0364	0.8	0.02986

Combi_2

	KNN	XGBoost	Random_Forest
Accuracy	0.90428	0.90428	0.91102
Precision	0.0	0.0	0.96825
Recall	0.0	0.0	0.07275
F1_Score	0.0	0.0	0.13533

Combi_3

	KNN	XGBoost	Random_Forest
Accuracy	0.97272	0.996	0.93761
Precision	0.0	1.0	0.22443
Recall	0.0	0.77918	0.99685
F1_Score	0.0	0.87589	0.36638

Combi_4

	KNN	XGBoost	Random_Forest
Accuracy	0.80925	0.83967	0.79603
Precision	0.12072	1.0	0.3966
Recall	0.02024	0.05261	0.39359
F1_Score	0.03466	0.09997	0.39509

Combi_5

	KNN	XGBoost	Random_Forest
Accuracy	0.95214	0.95214	0.95671
Precision	0.0	0.0	0.97619
Recall	0.0	0.0	0.09779
F1_Score	0.0	0.0	0.17778

Combi_6

	KNN	XGBoost	Random_Forest
Accuracy	0.98136	0.99934	0.67003
Precision	0.0	1.0	0.02661
Recall	0.0	0.92744	0.99685
F1_Score	0.0	0.96236	0.05183

Combi_7

	KNN	XGBoost	Random_Forest
Accuracy	0.79789	0.83028	0.78818
Precision	0.10221	0.0	0.33922
Recall	0.02496	0.0	0.26543
F1_Score	0.04012	0.0	0.29782

Combi_8

	KNN	XGBoost	Random_Forest
Accuracy	0.95214	0.95214	0.95331
Precision	0.0	0.0	0.62577
Recall	0.0	0.0	0.06082
F1_Score	0.0	0.0	0.11087

Combi_9

	KNN	XGBoost	Random_Forest
Accuracy	0.94558	0.99349	0.71986
Precision	0.0	1.0	0.03128
Recall	0.0	0.28076	1.0
F1_Score	0.0	0.43842	0.06067

For comparative analysis, a summary chart comparing the metrics across all 12 training/testing configurations of each method is presented in Figure 2, highlighting the best-performed metrics. For a more detailed visualization, Figures 3 and 4 provide focused insights into accuracy and F1-score trends, respectively. The F1-score is emphasized due to its relevance in leak detection, where both false alarms (precision) and missed leaks (recall) carry operational consequences. **F1-score** is emphasized due to the operational need to balance false alarms (costly inspections) and missed leaks (water loss) and the **Confusion matrices** highlight differences between models.

Overall, the best performance was observed in combination Combi_12, which achieved a precision of 0.9968 and an F1-score of 0.8, and in combination Combi_6, which reported an F1-score of 0.9624, both using XGBoost. However, several models exhibited either precision or recall values close to 0, suggesting possible data imbalance or overfitting issues. It is also noteworthy that Combi_3 achieved perfect precision (1.0) along with a decent recall (0.7792), whereas Combi_9 showed perfect recall (1.0) but very low precision (0.0313), both results being obtained with the Random Forest algorithm.

Figura 3 – Comparison of accuracy by technique and combination

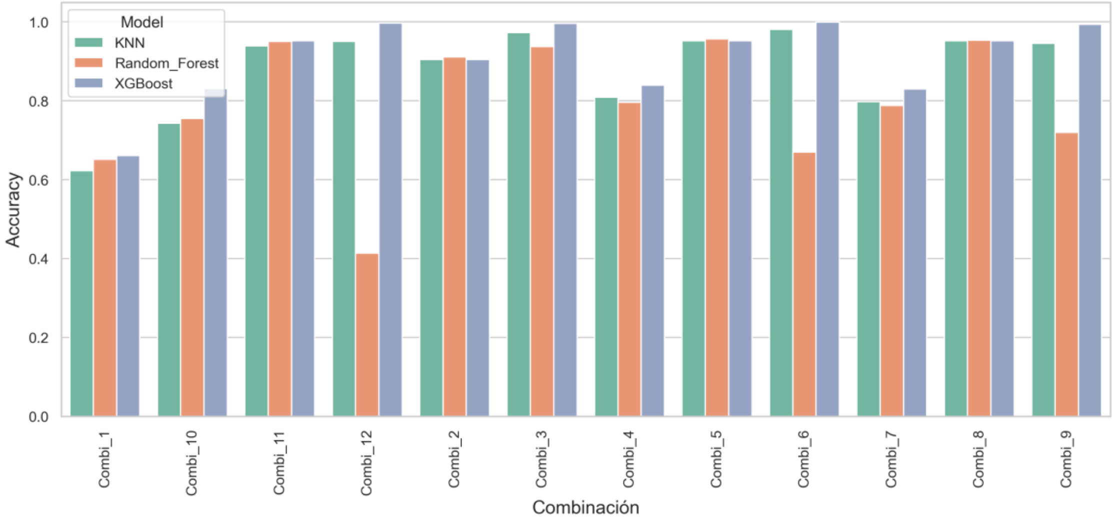
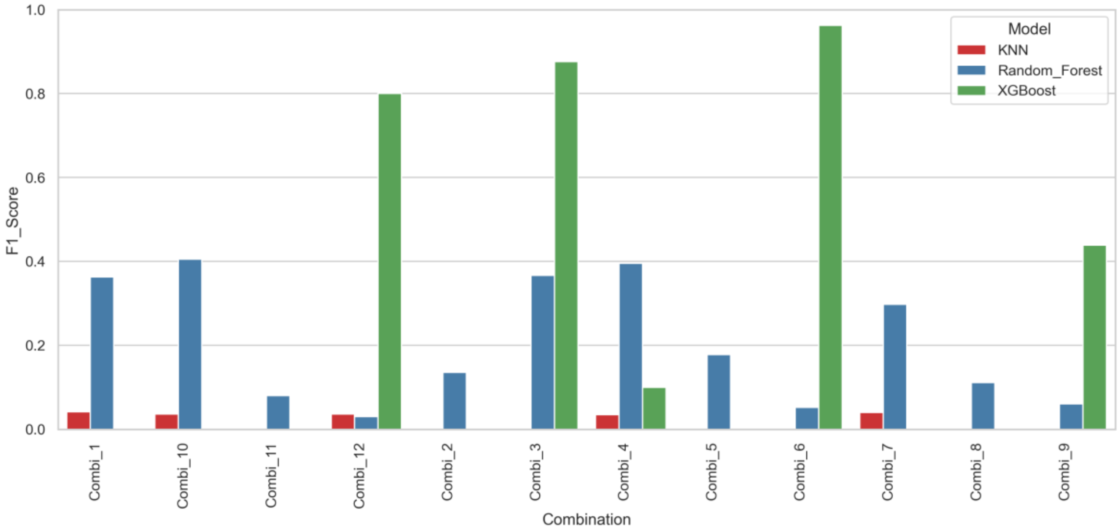


Figura 4 – Comparison of F1_Score by technique and combination



COMPARISION

The performance differences observed among the models can be attributed to how each algorithm handles the characteristics of the data, particularly class imbalance and the complexity of the data in the hydraulic simulations.

- XGBoost: Consistently outperformed the other algorithms across multiple combinations, including Combi_6 and Combi_12, showing both high precision and balanced F1-scores.
- Random Forest: Performed well in several combinations, especially in Combi_11, achieving a strong F1-score of 0.8759, but also suffered from inconsistencies.
- K-Nearest Neighbors (KNN): While simple and easy to implement, it struggled in most scenarios. It demonstrated lower performance compared to the other models, although it achieved acceptable results in a few specific configurations.

CONCLUSIONS AND FUTURE WORK

This study demonstrates the applicability of AI and ML techniques for binary leak detection in water distribution systems. Among the evaluated models, XGBoost emerged as the most consistent algorithm, suggesting its ability to capture the most informative features within the simulated hydraulic data. Random Forest also performed well, but exhibited some limitations, such as missing actual leaks or generating false positives in certain scenarios. K-NN showed the weakest performance overall, largely due to the high-class imbalance in the data, which hindered its ability to reliably distinguish leak events. These results reinforce the complexity of the problem being addressed.

These findings highlight the importance of selecting algorithms capable of handling complex, imbalanced datasets, particularly in critical infrastructure applications.

Future work will include:

- ✓ A deeper and refined analysis of the combinations, especially: Combi_6 The best all-rounder → F1 = 0.96, Precision = 1.0, Recall = 0.927, Combi_11: Very good balance between recall and precision, Combi_12: High precision with good recall. All with XGBoost, which may have the most informative features
- ✓ Expanding the analysis to multiclass classification for leak localization.
- ✓ Exploring how synthetic data could be closer to real-world data to improve training
- ✓ Incorporating real-world data to validate and improve the generalizability of models.
- ✓ Exploring additional AI techniques, such as deep learning and hybrid models.

ACKNOWLEDGMENTS

The author would like to thank the open-source community for providing the tools and libraries that made this work possible, and the Python scientific community for their continuous contributions to accessible and high-quality computational resources. Special appreciation is extended to the developers of WNTR and EPyT. This work was partially supported by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) – Brazil. Funding codes: [UNIVERSAL 4408164/2021-2], PDE [200014/2025-0].

REFERÊNCIAS

CHEN, T.; GUESTRIN, C. (2016). “XGBoost: A Scalable Tree Boosting System”. In *Anais do 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

COVER, T. M.; HART, P. E. (1967). “Nearest Neighbor Pattern Classification”. *IEEE Transactions on Information Theory*, 13(1), pp. 21–27.

POWERS, D.M.W. (2011). “Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation”. In Journal of Machine Learning Technologies, 2(1), pp. 37–63.

REIMAN, L. (2001). “Random Forests”. *Machine Learning*, 45(1), pp. 5–32.

SCHRÖER, C.; KRUSE, F.; GÓMEZ J. M. “A systematic literature review on applying CRISP-DM process model” *Procedia Computer Science*, 2021, Vol 181 pp. 526-534

SNIS (2018). *Perdas de água 2020: desafios para disponibilidade hídrica e avanço da eficiência do saneamento básico*. Trata Brasil, Water.org, GO Associados, São Paulo, junho 2020.

SNSA, SNIS (2023). Diagnóstico Temático Serviços de Água e Esgoto; Visão Geral ano de referência 2022. Ministério das Cidades Secretaria Nacional de Saneamento Ambiental, Sistema Nacional de Informações sobre Saneamento. Brasília, dezembro de 2023

STELIOS G. VRACHIMIS; MARIOS S. KYRIAKOU; DEMETRIOS G. ELIADES; MARIOS M. POLYCAPOU; “LeakDB: A benchmark dataset for leakage diagnosis in water distribution networks” *1st International WDSA / CCWI 2018 Joint Conference,* Kingston, Ontario, Canada July 23-25, 2018.