

# Eles aprendem diferente? Investigando padrões de aprendizado em Modelos Federados Globais e Centralizados com técnicas de Explainable AI (XAI)

Antonio Borssato<sup>1</sup>, Giovanni Comarella<sup>1</sup>

<sup>1</sup>Universidade Federal do Espírito Santo (UFES)  
Vitória – ES – Brasil

antonio.borssato@edu.ufes.br, gc@inf.ufes.br

**Abstract.** This work aims to answer the following question: do federated and centralized models learn the same decision patterns? To investigate this, a controlled scenario was created in which MLP and XGBoost models were trained both centrally and in a federated manner, and subsequently subjected to the same explainability evaluation process using the SHAP and LIME techniques. The results show that the federated models maintained performance competitive with the centralized ones and learned similar decision patterns.

**Resumo.** Este trabalho tem o objetivo de responder à seguinte pergunta: modelos federados e centralizados aprendem os mesmos padrões de decisão? Para isso, foi criado um cenário controlado onde modelos MLP e XGBoost foram treinados de forma centralizada e federada, e depois passaram pelo mesmo processo de avaliação de explicabilidade utilizando as técnicas SHAP e LIME. Os resultados mostraram que os modelos FL mantiveram performance competitiva aos centralizados e aprenderam padrões de decisão semelhantes.

## 1. Introdução

O aprendizado de máquina (ML) tem transformado diversas áreas nas últimas décadas, como saúde, setor bancário e indústria, ao possibilitar análise de grandes volumes de dados e automação de processos complexos [Ryan et al. 2025, Pugliese et al. 2021, Sarker 2021]. No seu formato clássico, o treinamento é considerado centralizado, onde todos os dados são armazenados num banco de dados único e o modelo aprende a partir desse conjunto. Entretanto, em áreas como a saúde, onde há a presença de dados sensíveis, a abordagem centralizada pode comprometer a privacidade. Para lidar com esse problema, surgiu a técnica de aprendizado federado (FL), que permite o treinamento colaborativo sem a troca direta dos dados [McMahan et al. 2023, Liu et al. 2023]. No FL, os participantes treinam modelos menores localmente nas suas partições de dados e comunicam apenas atualizações de pesos ou gradientes com um servidor global, responsável por agregar um modelo final, reduzindo a exposição dos dados e, em muitos casos, a necessidade de transferência intensa de informação [Wang 2019].

Mesmo que o FL garanta a privacidade dos dados compartilhados, ainda nos deparamos com um questionamento clássico do aprendizado de máquina: como esses modelos tomam decisões? Responder essa pergunta é particularmente importante para os domínios onde os usuários precisam entender o que influencia uma predição do modelo, como na saúde [Ribeiro et al. 2016, Salih et al. 2024]. Métodos de Inteligência Artificial

Explicável (XAI) como SHAP (baseado em valores de Shapley) e LIME (aproximação local por modelos lineares) são comumente usados para compreender modelos ao medir a importância das features e fornecer explicações locais e globais do comportamento do modelo [Lundberg e Lee 2017, Amoroso 2023].

Diante da questão de explicabilidade tanto dos modelos centralizados quanto federado, o presente trabalho busca responder a seguinte pergunta: modelos federados e centralizados aprendem os mesmos padrões para tomada de decisões? Para investigar essa questão, este trabalho propõe um cenário controlado, usando as mesmas configurações e conjunto de dados de precipitação na Austrália e equivalendo, na medida do possível, pré-processamento, particionamento e hiperparâmetros, de modo que as diferenças sejam atribuídas somente à arquitetura de treinamento (centralizada versus federada). É importante ressaltar que algumas decisões do cenário controlado, como pré-processamento global antes da repartição violam premissas do FL. Essa escolha é deliberada para garantir comparabilidade experimental entre as arquiteturas. Em outros trabalhos, como [Rumesh et al. 2023], é explorado um cenário mais realista de FL para comparação com modelos centralizados e avaliação de XAI. Eles concluem que as explicações SHAP fornecidas pelo modelo federado diferem das do modelo centralizado à medida que a porcentagem de anomalias nos dados locais do cliente aumenta e torna as partições mais heterogêneas.

Para a execução desse trabalho foram consideradas duas arquiteturas de modelos: um MLP (rede neural do tipo feedforward) e XGBoost (ensemble de árvores baseado em boosting) [Taud e Mas 2018, Niazkar et al. 2024]. No contexto federado, o MLP foi treinado com o agregador FedAvg (média ponderada das atualizações dos clientes) e o XGBoost foi adaptado com duas estratégias de agregações: Bagging e Cyclic [Claude Sammut 2010, Jiang et al. 2024]. Em todos os cenários, foram aplicados SHAP e LIME de forma idêntica para comparar a importância das features produzida por cada arquitetura e modelo e entender seus padrões de decisão.

## 2. Metodologia

### 2.1. Conjunto de Dados

O estudo utilizou o dataset Rain in Australia [Young e Young 2020], que contém cerca de 10 anos de observações de 44 estações meteorológicas. O conjunto original possui 145.460 registros e 23 variáveis. O pré-processamento removeu atributos com mais de 35% de dados ausentes (Sunshine, Evaporation, Cloud3pm, Cloud9am) e a variável Date, que era irrelevante para a tarefa. As instâncias restantes com valores nulos foram descartadas, resultando em 112.925 registros. A variável alvo RainTomorrow é um indicador booleano se choverá amanhã e apresenta desbalanceamento (22.16% da classe Chuva e 77.84% da classe Não Chuva). Para simular o cenário federado, os dados foram divididos por estação meteorológica (Location). Além disso, a análise de heterogeneidade dos dados em cada um dos 44 clientes via Divergência de Jensen-Shannon [G. et al. 2024] indicou distribuição Non-IID moderada (média 0.0483), o que significa que cada partição de dados não foge tanto a distribuição total do conjunto [Wei et al. 2024].

### 2.2. Divisão e validação dos dados

O dataset foi dividido de forma estratificada em treino (70%), validação (10%) e teste (20%), preservando a proporção de classes e distribuição de localidade. Para as variáveis

categóricas cíclicas de direção do vento, aplicou-se codificação trigonométrica (seno e cosseno) para preservar o sentido angular da rosa dos ventos, totalizando 19 features de entrada. Quanto à normalização dos dados, somente foi executada para os modelos MLPs, onde Z-score foi aplicada globalmente para o cenário centralizado e localmente para cada um dos 44 clientes no cenário federado. Além disso, para as análises de explicabilidade, 2000 amostras foram selecionadas aleatoriamente no conjunto de teste e utilizadas em todos os modelos. Esse valor foi definido considerando as limitações de poder computacional e também é estratificado respeitando a distribuição da variável alvo e location.

### 2.3. Algoritmos e modelos

Foram comparadas duas arquiteturas: (1) MLP (Multilayer Perceptron) [Goodfellow et al. 2016]: composta por três camadas ocultas com ativação ReLU e Dropout, otimizada via Adam [Kingma e Ba 2017] e perda Binary Cross-Entropy; (2) XGBoost: ensemble de árvores [Chen e Guestrin 2016] com otimização de log-loss e métrica AUCPR.

Os hiperparâmetros foram otimizados via otimização bayesiana com Optuna [Akiba et al. 2019] no cenário centralizado e replicados para o federado, para que os parâmetros fossem equivalentes nas duas arquiteturas de treinamento. Para isso, foram executados 60 *trials* para MLP e 200 para XGBoost. Após a otimização, realizou-se o treinamento final centralizado com 150 épocas (*early stopping* com paciência de 15 épocas) para MLP e 500 rounds (*early stopping* com paciência de 20 rounds) para o XGBoost, monitorando a AUCPR de validação.

No contexto federado, implementado via framework Flower [Beutel et al. 2022], utilizou-se: (1) MLP FedAvg: Agregação de pesos por média ponderada [McMahan et al. 2023] após treinamento local de 5 épocas por round (30 rounds totais); e (2) XGBoost Federado: testou-se a estratégia Bagging (treinamento paralelo dos 44 clientes em um único round, com 5 árvores cada, agregadas por média dos modelos) [Claude Sammut 2010] e Cyclic (treinamento sequencial onde o modelo de 5 árvores de um cliente é ponto de partida para o próximo, que corrige os erros do anterior) [Jiang et al. 2024], ambas gerando um modelo global de 220 árvores.

### 2.4. Avaliação do Treinamento e Comparações de Explicabilidade

O desempenho dos modelos foi avaliado primordialmente pela métrica **AUCPR** (Área sob a curva Precision–Recall, também conhecida como Average Precision), por ser mais adequada para datasets desbalanceados, já que foca na performance da classe minoritária. Outra métricas auxiliares foram calculadas, que incluem F1-score, Precision, Recall e Matthews Correlation Coefficient (MCC), que é bom para dados desbalanceados [Chicco e Jurman 2020], além da matriz de confusão.

A explicabilidade foi avaliada em 2000 amostras fixas do conjunto de teste. Para o SHAP, utilizou-se o GradientExplainer (MLP) com 100 amostras de background e TreeExplainer (XGBoost) para calcular valores exatos ou aproximados de contribuição [Lundberg e Lee 2017]. Para o LIME, foram geradas 500 perturbações por instância para ajustar o modelo linear local [Ribeiro et al. 2016]. As importâncias das features foram obtidas como a média dos valores absolutos de SHAP na classe de Chuva e, da mesma forma, pela média dos módulos dos coeficientes do surrogate linear do LIME.

As features foram ranqueadas por importância para a comparação entre modelos. Além disso, a fidelidade do LIME foi avaliada pela distribuição de  $R^2$  entre as previsões do surrogado e as previsões do modelo global, sendo  $R^2 > 0,80$  considerado boa aproximação local [Gaudel et al. 2022]. Como forma de análise qualitativa, foram empregados os gráficos de beeswarm plots de SHAP para inspecionar direção (efeito positivo ou negativo), magnitude e não linearidades das contribuições das features.

### 3. Resultados e Discussão

#### 3.1. Otimização de Hiperparâmetros

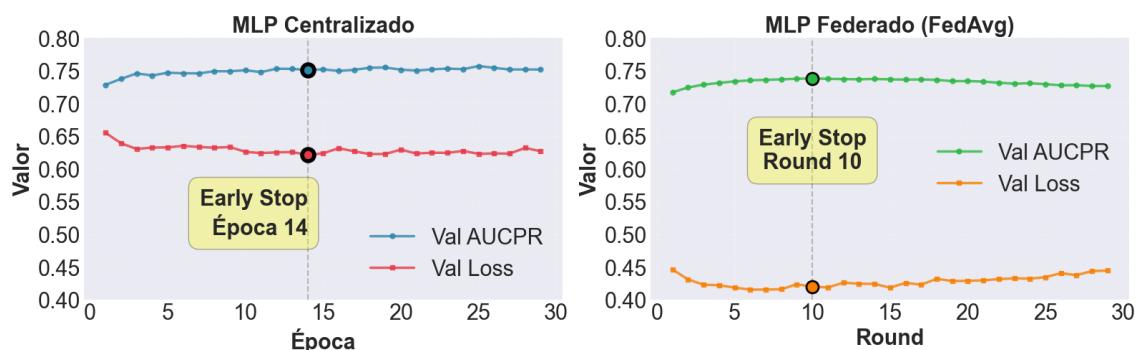
A otimização Bayesiana identificou as configurações ótimas resumidas na Tabela 1. Como descrito no processo de metodologia da Subseção 2.3, esses foram os parâmetros utilizados tanto nos modelos finais centralizados quanto federados.

**Tabela 1. Hiperparâmetros ótimos identificados via otimização Bayesiana.**

Hiperparâmetro (MLP)	Valor	Hiperparâmetro (XGBoost)	Valor
Hidden layer 1	256	eta (learning rate)	0.0398532599
Hidden layer 2	80	max_depth	6
Hidden layer 3	64	min_child_weight	1
Dropout 1	0.1587722675	gamma	0.3997136853
Dropout 2	0.1990458090	subsample	0.7236119531
Learning rate	0.0014015879	colsample_bytree	0.8784668317
Batch size	32	reg_alpha	0.0003571243
scale_pos_weight	3.5138761992	scale_pos_weight	3.5138761992
		reg_lambda	0.0034469898

#### 3.2. Análise de Convergência no treinamento

A Figura 1 apresenta curvas de aprendizado para MLP Centralizado e Federado.

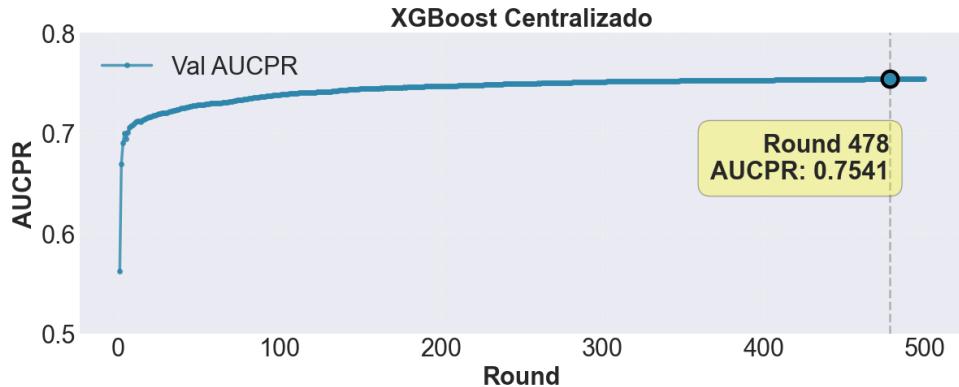


**Figura 1. Curvas de aprendizado MLP centralizado e federado.**

O MLP Centralizado convergiu em 29 épocas com early stopping (paciência = 15), atingindo AUCPR de validação de 0.7510 e loss de 0.6210 na época 14 (melhor modelo salvo). O modelo federado convergiu em 10 rounds de comunicação (30 rounds totais), com AUCPR de validação agregado de 0.7382. Observa-se que ambos os modelos

apresentam pouca oscilação nas métricas de validação AUCPR, o que mostra treinamentos estáveis e valida a estratégia de agregação FedAvg. Percebe-se ainda que o modelo federado possui menor loss, em todo o treinamento, mas alcançou um AUCPR levemente menor.

A Figura 2 apresenta curvas de aprendizado para o XGBoost.



**Figura 2. Curvas de aprendizado XGBoost centralizado.**

Para o XGBoost Centralizado, o treinamento utilizou 499 rounds de boosting com early stopping de 20 rounds, portanto, o melhor round foi o 478, onde alcançou 0.7541 de AUCPR de validação. Nos modelos federados, a estratégia Bagging utilizou 220 árvores divididas entre 44 clientes (5 árvores/cliente) em um único round de agregação, enquanto a Cyclic utilizou 220 árvores ao longo de 44 rounds sequenciais (5 árvores/round). A ausência de métricas de convergência intermediárias nos XGBoosts federados reflete limitação arquitetural da implementação utilizada no Flower, onde são registradas apenas as métricas do modelo final. A estratégia Bagging alcançou AUCPR de 0.69 (0.10), enquanto a Cyclic alcançou 0.70 (0.10).

### 3.3. Performance Preditiva

A Tabela 2 apresenta as métricas de performance dos cinco modelos avaliados no conjunto de teste (22.585 amostras, 20% do dataset).

**Tabela 2. Performance dos modelos no conjunto de teste.**

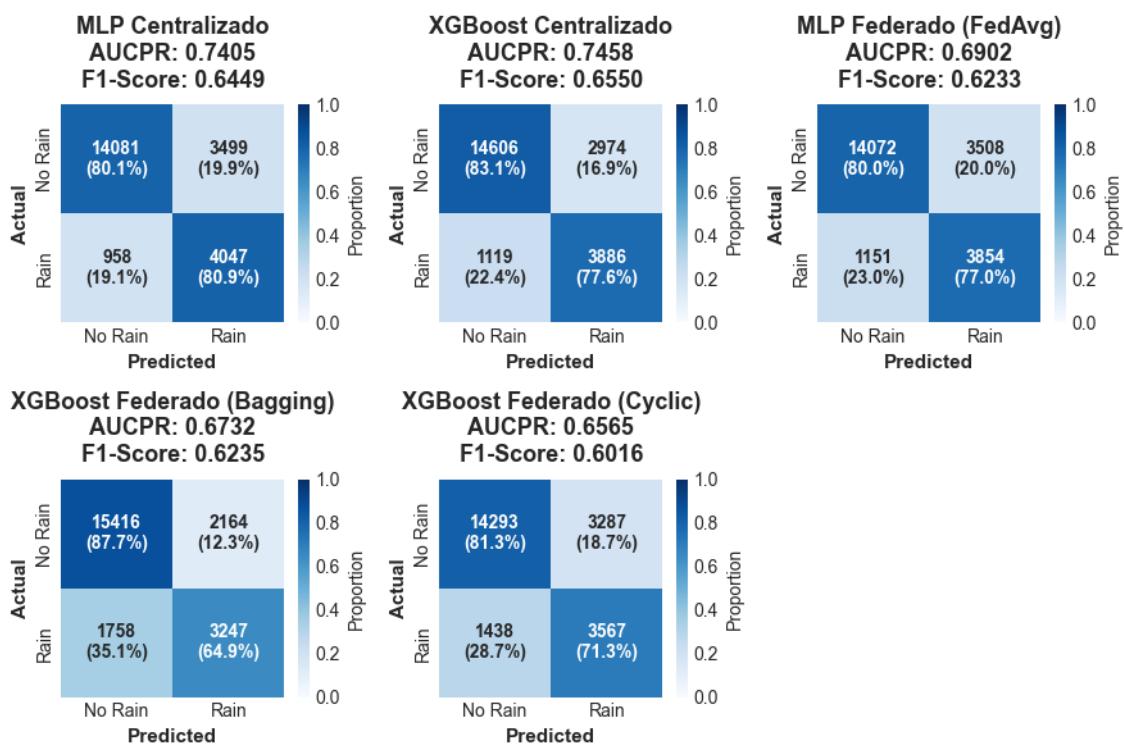
Modelo	AUCPR	Precision	Recall	F1-Score	MCC
XGBoost Centralizado	0.7458	0.5665	0.7764	0.6550	0.5484
MLP Centralizado	0.7405	0.5363	0.8086	0.6449	0.5367
MLP Federado (FedAvg)	0.6904	0.5221	0.7728	0.6232	0.5055
XGBoost Fed. (Bagging)	0.6732	0.6001	0.6488	0.6235	0.5115
XGBoost Fed. (Cyclic)	0.6565	0.5204	0.7127	0.6016	0.4749

Os modelos centralizados alcançaram performance superior - o que vai ao encontro dos achados da literatura [Rumesh et al. 2023] - com XGBoost Centralizado obtendo AUCPR de 0.7458 (melhor resultado global) e MLP Centralizado 0.7405, diferença de apenas 0.71%. Ambos demonstram capacidade preditiva acima do baseline aleatório (AUCPR = 0.222, equivalente à proporção da classe minoritária). O MLP Centralizado

apresentou maior recall (0.8086), enquanto o XGBoost Centralizado equilibrou precision (0.5665) e recall (0.7764), resultando no maior F1-Score.

Entre os modelos federados, o MLP Federado (FedAvg) obteve melhor desempenho com AUCPR de 0.6904, representando degradação de 6.77% em relação ao centralizado. Percebe-se então um *trade-off* entre privacidade e performance, onde cada cliente possui distribuições distintas de padrões climáticos, mesmo com Non-IDD moderado. As estratégias XGBoost federadas apresentaram degradação mais acentuada: Bagging com 9.73% (AUCPR 0.6732) e Cyclic com 11.97% (AUCPR 0.6565). A estratégia Bagging, que agrupa modelos de todos os 44 clientes em paralelo em um único round, demonstrou melhor precision (0.6001) entre os federados, mas recall inferior (0.6488). A estratégia Cyclic, que treina sequencialmente um cliente por round ao longo de 44 rounds, apresentou o menor AUCPR global, mas manteve recall competitivo (0.7127), sugerindo que a exposição sequencial a diferentes distribuições locais de dados diminui a capacidade de generalização global.

A Figura 3 apresenta as matrizes de confusão normalizadas por linha (recall) para os cinco modelos.



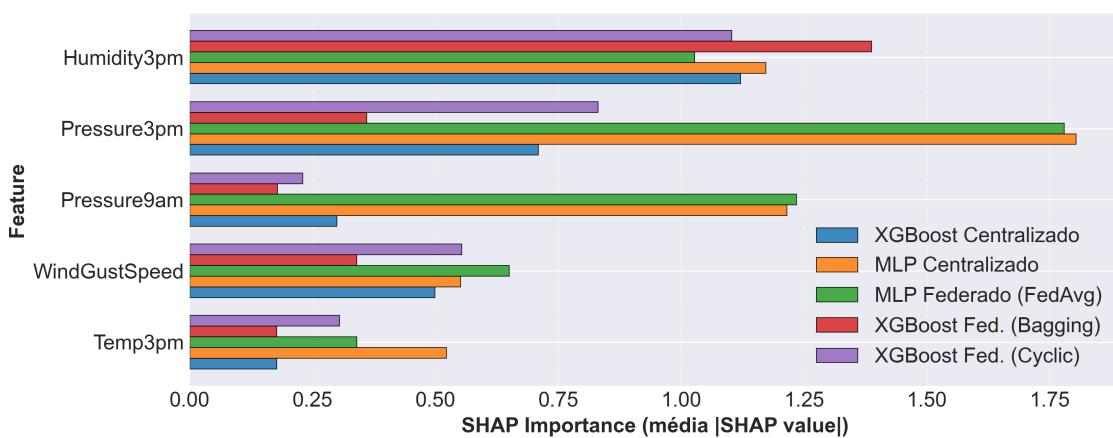
**Figura 3. Matrizes de Confusão - Normalizado por Linha (Recall)**

Observa-se que os modelos têm tendência de falsos positivos considerável. O MLP Federado mantém padrão similar ao XGBoost centralizado (aproximadamente 77% recall para chuva), enquanto os XGBoosts federados priorizam especificidade: o Bagging atinge 87.7% de classificações corretas para Não Chuva, mas detecta apenas 64.9% dos eventos de chuva, e o Cyclic apresenta maior balanceamento de recall (81.4% No Rain, 71.3% Rain).

A hierarquia de performance XGBoost Centralizado > MLP Centralizado > MLP Federado > XGBoost Bagging > XGBoost Cyclic mantém-se consistente em AUCPR, F1-Score e MCC. A métrica MCC, que varia de -1 (predições inversas) a +1 (perfeitas), com 0 indicando predição aleatória, mostra valores entre 0.47 e 0.55, que confirmam uma capacidade de predição correta considerável nos modelos.

### 3.4. Explicabilidade via SHAP

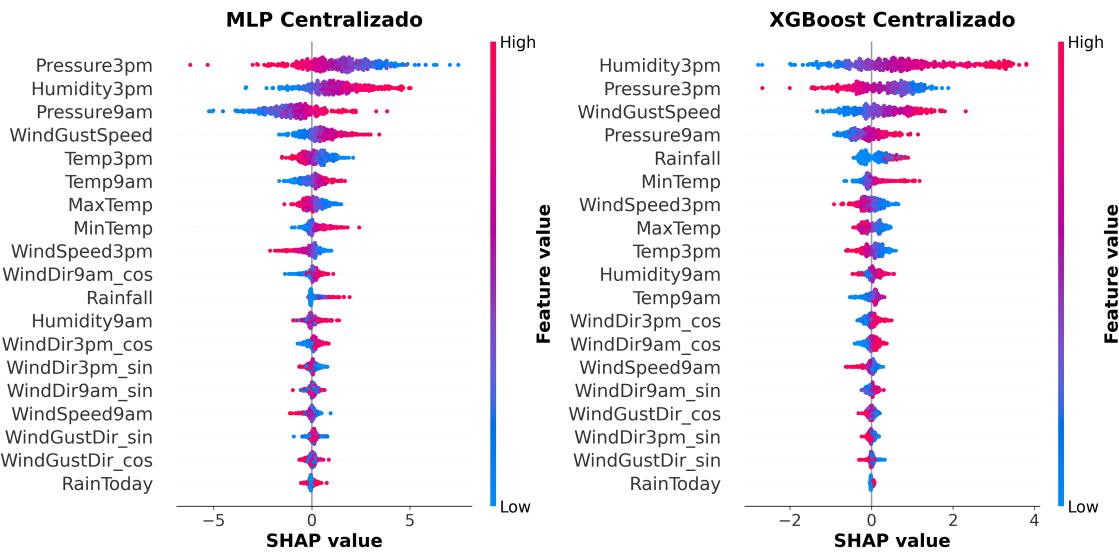
A análise de explicabilidade global via SHAP identificou as features mais relevantes para cada modelo, apresentadas na Figura 4.



**Figura 4. Top 5 Features por Importância SHAP para a classe Chuva**

Para os modelos centralizados MLP e XGBoost, e os modelos federados MLP e XGBoost bagging, as quatro features mais importantes são consistentemente: Humidity3pm, Pressure3pm, Pressure9am e WindGustSpeed. No Caso da estratégia cyclic, apenas a Pressure9am que não está no top 4, que é ocupado pela MaxTemp. A concordância entre MLP e XGBoost nestas features sugere captura de relações causais meteorológicas verdadeiras, não apenas artefatos do modelo. Percebe-se também que as abordagens federadas e centralizadas consideram as mesmas features como as mais importantes, com pequenas alterações nas ordens. Como é observado em [Rumesh et al. 2023], um dos fatores de discrepância nas explicações federado vs centralizado é a alta heterogeneidade dos dados entre os clientes, que não é o caso desse trabalho. Entretanto, a diferença nas magnitudes das importâncias sugere que o aprendizado federado dilui ligeiramente os padrões aprendidos de algumas features. Em relação às agregações do XGBoost federado, percebe-se que o bagging se apoia fortemente na umidade das 15h, enquanto o cyclic mostra um padrão mais uniforme e generalista.

Além da importância de features, as distribuições de contribuições individuais (beeswarms) para a classe Rain encontram-se nas Figuras 5–6. De modo geral, observa-se uma alta concordância entre os modelos, onde valores maiores e menores das features direcionam as predições dos cinco modelos para uma mesma decisão. Nota-se, portanto, que valores maiores para umidade às 15h, pressão às 9h e a velocidade da rajada de vento nas 24h até a meia noite, junto de valores menores para a pressão às 15h, tendem a gerar predições de chuva em todos os modelos.



**Figura 5. SHAP Beeswarm dos Modelos Centralizados - Classe Rain**

Comparando as arquiteturas de modelos, os MLP apresentam uma amplitude de valores SHAP maior que os XGBoosts (dispersão horizontal maior nos beeswarms), sugerindo que o MLP é mais sensível a variações de algumas features em amostras individuais. Já o XGBoost tende a distribuir a importância de forma mais concentrada, com magnitudes absolutas geralmente menores.

Em resumo, os beeswarms confirmam que as previsões estão baseadas em variáveis meteorológicas plausíveis do ponto de vista físico (umidade, pressão e rajada de vento), que a ordenação relativa das features é consistente entre modelos, e que as abordagens federadas preservam esse ordenamento apesar de suavizarem a intensidade das importâncias.

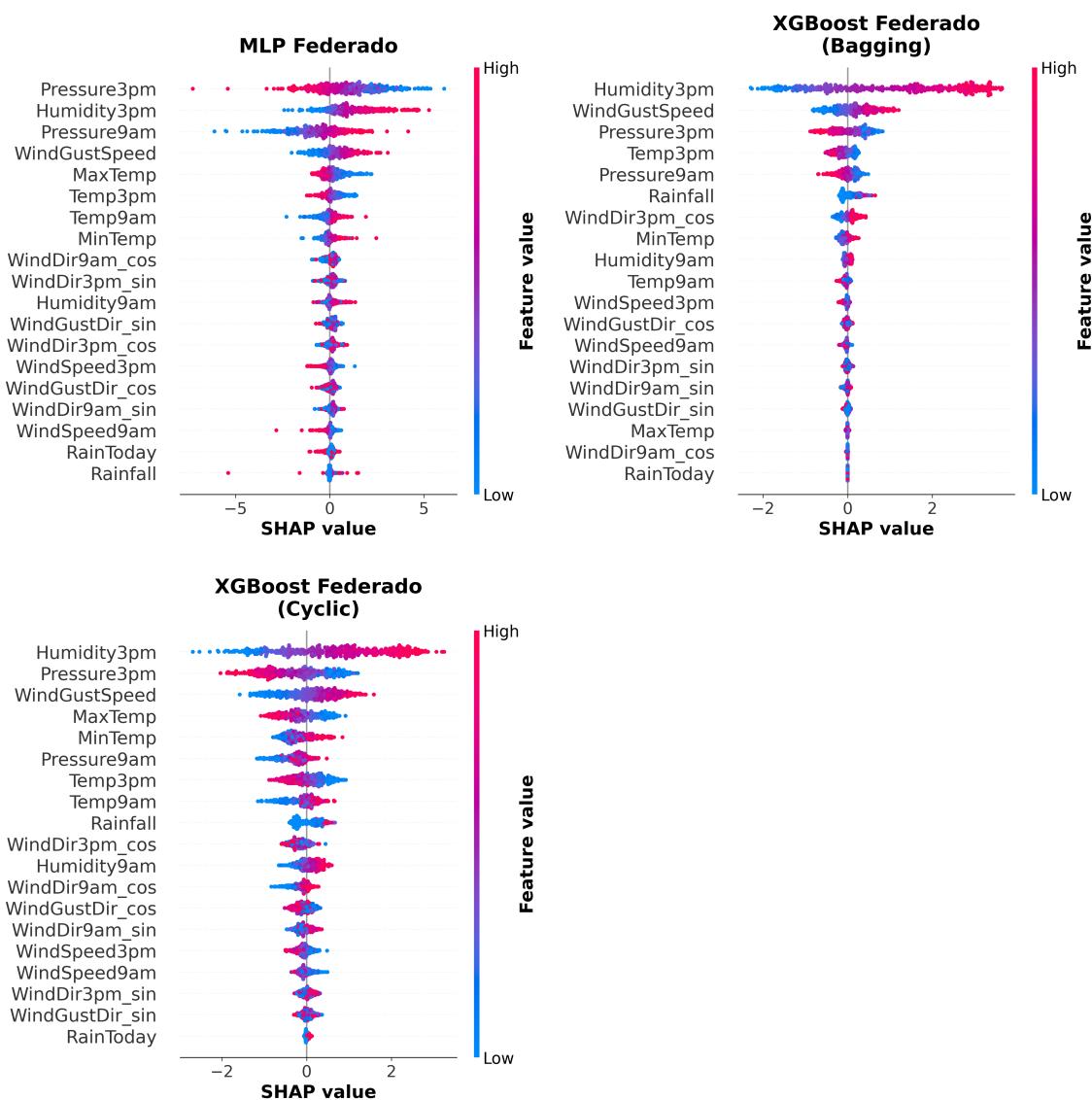
### 3.5. Explicabilidade via LIME

A análise LIME complementa os resultados SHAP ao fornecer explicações locais individuais. A Figura 7 apresenta a importância das cinco features mais relevantes para os modelos.

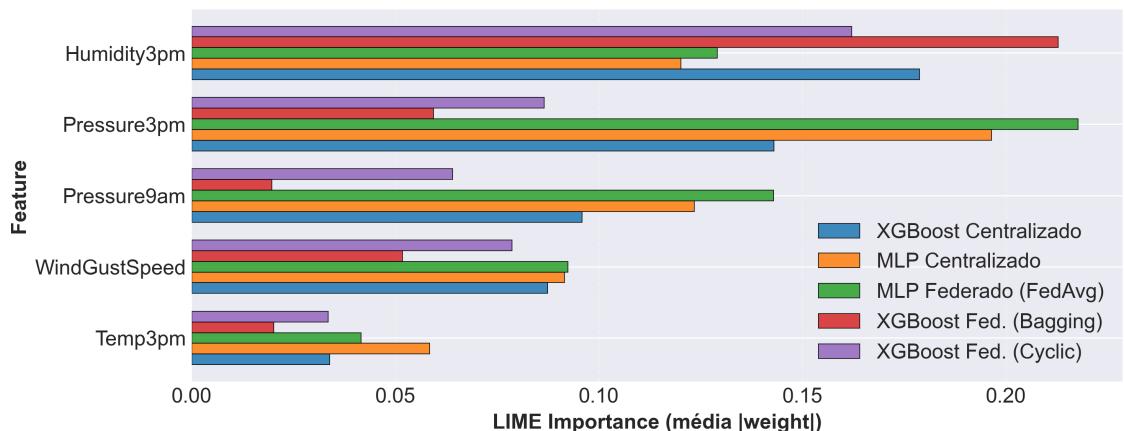
Percebe-se que os resultados do LIME estão alinhados àqueles encontrados via SHAP. Conforme ilustrado na Figura 7, o conjunto das cinco features mais importantes Humidity3pm, Pressure3pm, Pressure9am, WindGustSpeed e Temp3pm é consistente entre as duas metodologias de explicabilidade. Para complementar a análise, podemos ver na Tabela 3 a fidelidade das aproximações locais do LIME aos modelos reais.

Os resultados de fidelidade indicam que o LIME é eficaz para aproximar o modelo linear aos modelos MLP ( $R^2 > 0,92$ ), o que garante boa interpretabilidade. Já para os modelos XGBoost, a fidelidade cai consideravelmente, situando-se entre 0,50 e 0,66, com destaque negativo para o cenário federado Cyclic. Essa baixa fidelidade nos modelos XGBoost sugere uma complexidade local e não-linearidade que dificultam a aproximação do LIME e pode ser um indicativo de baixa estabilidade das explicações para essa arquitetura baseada em árvores.

Uma observação relevante ao comparar as duas técnicas de XAI é que, nas Figu-



**Figura 6. SHAP Beeswarm dos Modelos federados - Classe Rain**



**Figura 7. Top 5 Features por Importância LIME para a classe Chuva**

**Tabela 3. Comparação da Fidelidade Local ( $R^2$  Score) do LIME entre os modelos.**

Modelo	$R^2$ Médio	Desvio Padrão
MLP Centralizado	0.9292	0.0773
MLP Federado (FedAvg)	0.9250	0.0778
XGBoost Centralizado	0.6652	0.0829
XGBoost Federado (Bagging)	0.6200	0.1069
XGBoost Federado (Cyclic)	0.5088	0.0831

ras 4–7, os modelos baseados em MLP tendem a concentrar suas maiores importâncias nas variáveis de pressão, enquanto os modelos XGBoost apresentam uma distribuição mais equilibrada entre diferentes features. Contudo, ao analisar as explicações fornecidas pelo LIME, observa-se uma mudança no comportamento do XGBoost: o modelo passa a atribuir importância substancialmente maior à variável de umidade às 15h. Em outras palavras, o LIME sugere que as redes neurais tendem a capturar padrões mais sensíveis à pressão, ao passo que o modelo baseado em árvores concentra maior atenção na umidade para fazer previsões.

Em relação às estratégias de agregação federada do XGBoost, os resultados também ecoam as observações do SHAP. O modelo XGBoost Fed. (Bagging) demonstra uma forte dependência de uma única feature, Humidity3pm (0.213), que se destaca como a mais importante. Em contraste, o XGBoost Fed. (Cyclic), embora ainda priorize Humidity3pm (0.162), distribui a importância de forma mais equilibrada entre outras features, como Pressure3pm (0.087) e WindGustSpeed (0.079), confirmando o padrão mais generalista dessa abordagem de agregação.

### 3.6. Síntese dos Resultados

Em suma, a análise LIME corrobora os achados do SHAP. A alta concordância na identificação das features mais relevantes (umidade e pressão) entre diferentes modelos e metodologias fortalece a conclusão de que os modelos estão capturando padrões nas relações meteorológicas de forma genuína. Mais importante, isso demonstra que os modelos federados, em cenário controlado e a heterogeneidade dos dados não tão elevada, aprendem a focar nas mesmas variáveis preditivas que seus equivalentes centralizados, validando sua eficácia para a tarefa. Além disso, é importante lembrar que o desempenho dos modelos federados foi um pouco inferior aos centralizados, o que também é visto em outros trabalhos.

Os resultados possibilitam responder as dúvidas apresentadas na introdução: os modelos centralizados alcançam performance superior devido ao acesso irrestrito aos dados completos; o aprendizado federado preserva privacidade, mas incorre em degradação de performance; técnicas de XAI (SHAP e LIME) mantêm-se interpretáveis em ambos os contextos, identificando as mesmas features críticas de umidade e pressão, embora com magnitudes diferentes nos modelos federados e menor fidelidade aos modelos de árvore. A estratégia MLP FedAvg demonstrou melhor equilíbrio entre privacidade e performance para os federados, enquanto XGBoost Cyclic apresentou o menor desempenho entre as estratégias de agregação de árvores.

## 4. Conclusão

O presente trabalho buscou investigar, por meio de técnicas de Inteligência Artificial Ex-  
plicável (XAI), se os modelos treinados em arquiteturas Federadas (FL) e Centralizadas  
aprendem os mesmos padrões de decisão. A análise comparativa em cenário controlado,  
que envolveu modelos MLP e XGBoost na tarefa de previsão de chuva na Austrália,  
demonstrou que, dadas as limitações do experimento e do cenário simulado, as duas abor-  
dagens convergem para uma mesma explicação dos padrões preditivos.

A aplicação de SHAP e LIME revelou consistência no ranqueamento das features  
mais importantes em todos os cinco modelos analisados. Variáveis como Humidity3pm,  
Pressure3pm, Pressure9am e WindGustSpeed foram identificadas de forma estável como  
os pilares de decisão.

Os resultados encontrados validam a eficácia do aprendizado federado. O FL,  
mesmo trabalhando com dados distribuídos com heterogeneidade moderada, foi capaz de  
aprender os mesmos padrões preditivos dos modelos centralizados. Esse comportamento  
foi especialmente evidente nas redes MLP, que apresentaram maior consistência entre as  
versões federada e centralizada do que o XGBoost.

## Referências

- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-  
generation hyperparameter optimization framework.
- Amoroso, F. S. (2023). Inteligência artificial explicável com lime e shap aplicada à rede  
neural convolucional. Trabalho de Conclusão de Curso (Graduação). Orientador: Pe-  
reira, Clayton Reginaldo.
- Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li,  
K. H., Parcollet, T., de Gusmão, P. P. B., and Lane, N. D. (2022). Flower: A friendly  
federated learning research framework.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proce-  
edings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery  
and Data Mining*, KDD '16, page 785–794. ACM.
- Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient  
(mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*,  
21.
- Claude Sammut, G. I. W., editor (2010). *Encyclopedia of Machine Learning*. Springer,  
New York, NY.
- G., D. M. J., Solans, D., Heikkila, M., Vitaletti, A., Kourtellis, N., Anagnostopoulos,  
A., and Chatzigiannakis, I. (2024). Non-iid data in federated learning: A survey with  
taxonomy, metrics, methods, frameworks and future directions.
- Gaudel, R., Galárraga, L., Delaunay, J., Rozé, L., and Bhargava, V. (2022). s-lime: Re-  
conciling locality and fidelity in linear explanations.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- Jiang, N., Huang, Z., and Sui, Y. (2024). Explanation-driven cyclic learning for high-  
quality brain mri reconstruction from unknown degradation. In *Medical Image Com-*

*puting and Computer Assisted Intervention – MICCAI 2024*, pages 318–328, Cham. Springer Nature Switzerland.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Liu, B., Lv, N., Guo, Y., and Li, Y. (2023). Recent advances on federated learning: A systematic survey.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2023). Communication-efficient learning of deep networks from decentralized data.

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., and Righetti, M. (2024). Applications of xgboost in water resources engineering: A systematic literature review (dec 2018–may 2023). *Environ. Model. Softw.*, 174(C).

Pugliese, R., Regondi, S., and Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Science and Management*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Rumesh, Y., Senevirathna, T. T., Porambage, P., Liyanage, M., and Yliantila, M. (2023). Comprehensive analysis over centralized and federated learning-based anomaly detection in networks with explainable ai (xai). In *ICC 2023 - IEEE International Conference on Communications*, pages 4853–4859.

Ryan, A., Pratama, Wabula, F., Ilmandry, H., Isabela, M. L., Raharjo, M., and Sianipar, R. (2025). Literature review the impact of machine learning in modern industries. *Nian Tana Sikka : Jurnal ilmiah Mahasiswa*.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., and Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, 7(1).

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *Sn Computer Science*, 2.

Taud, H. and Mas, J. (2018). Multilayer perceptron (mlp).

Wang, G. (2019). Interpret federated learning with shapley values.

Wei, B., Li, J., Liu, Y., and Wang, W. (2024). Non-iid federated learning with sharper risk bound. volume 35, pages 6906–6917. IEEE. Epub 2024 May 2. PMID: 36279343.

Young, J. and Young, A. (2020). Rain in australia: Predict next-day rain in australia. <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package>. Acesso em: 10 nov. 2025.