Do Aprendizado Centralizado ao Federado: O Que Acontece com a Explicabilidade dos Modelos?

Daniel Ribeiro Trindade[1,2], **Eduardo Zambon**[1], **Rodolfo da Silva Villaça**[1]
Diego Roberto Colombo Dias[1], **Giovanni Comarela**[1]

[1] Universidade Federal do Espírito Santo (Ufes)
{rodolfo.villaca, diego.dias}@ufes.br, {zambon, gc}@inf.ufes.br

[2]Instituto Federal do Espírito Santo (Ifes)
danielrt@ifes.edu.br

Abstract. *This paper investigates the impact of federated learning (FL) on the explainability of machine learning models, with a focus on SHAP-based explanations. A comparative methodology is proposed between centrally trained and federated models, considering scenarios with uniform and non-uniform data distributions across clients. The approach evaluates the similarity of explanations using metrics such as cosine distance, ranking similarity, and sign consistency, and is validated using the EHMS dataset for attack detection in healthcare systems. Results show that federated models produce explanations that differ from those obtained with centralized training, and data heterogeneity significantly affects the explanation consistency: in uniform scenarios, local and global explanations are consistent, whereas in non-uniform settings they diverge.*

Resumo. *Este artigo investiga o impacto do aprendizado federado (FL) na explicabilidade de modelos de aprendizado de máquina, com foco em explicações baseadas em valores SHAP. É proposta uma metodologia comparativa entre modelos treinados de forma centralizada e federada, considerando cenários com distribuição uniforme e não uniforme dos dados entre clientes. A abordagem avalia a similaridade das explicações por meio de métricas como distância do cosseno, similaridade de rankings e consistência de sinais, sendo validada com o conjunto de dados EHMS para detecção de ataques em sistemas de saúde. Os resultados mostram que modelos federados produzem explicações distintas das centralizadas e que a heterogeneidade dos dados afeta significativamente a consistência das explicações. Em cenários uniformes, explicações locais e globais são consistentes, enquanto em cenários não uniformes tornam-se divergentes.*

1. Introdução

O uso de modelos de Inteligência Artificial (IA) é cada vez mais comum em diversas áreas. Setores sensíveis, tais como a segurança de redes e sistemas de saúde [Hady et al. 2020], têm se beneficiado bastante com a aplicação da IA. Essa ampla adoção, entretanto, gera também questionamentos sobre como esses modelos operam, além de aspectos éticos e de privacidade de dados [Branson et al. 2020, Yang et al. 2020, Bak et al. 2024]. Para atacar esses problemas, diversas abordagens podem ser seguidas. Este trabalho foca em duas: IA explicável (XAI – *eXplainable Artificial Intelligence*) e Aprendizado Federado (FL – *Federated Learning*).

O FL aborda um problema importante em situações nas quais os dados não estão centralizados [McMahan et al. 2023], ou não podem ser compartilhados [Yang et al. 2020]. Em vez de enviar os dados para um servidor central, o FL permite que o modelo de IA seja treinado diretamente nos dispositivos locais, sem a necessidade de compartilhamento dos dados em si. Por sua vez, XAI se refere ao desenvolvimento de técnicas que possam ajudar a explicar e interpretar as decisões dos modelos de IA de forma compreensível para seres humanos [Linardatos et al. 2021]. Isso ajuda a aumentar a confiança nas decisões da IA, pois permite que usuários entendam o raciocínio por trás das respostas ou ações dos sistemas [Markus et al. 2021]. Dentre as técnicas de interpretabilidade, o SHAP (*SHapley Additive exPlanations*) [Lundberg and Lee 2017] destaca-se pela sua robustez teórica, fundamentada na Teoria dos Jogos. Diferente de outros métodos agnósticos ao modelo, o SHAP garante propriedades matemáticas de equidade na distribuição das contribuições de cada característica (*feature*), independentemente da arquitetura do modelo utilizado.

Em separado, XAI e FL possuem uma vasta quantidade de pesquisa consolidada, mas há poucos estudos sobre o uso dessas abordagens de forma integrada [Lopez-Ramos et al. 2024]. Por exemplo, alguns trabalhos fazem uso de XAI para influenciar no treinamento de modelos em ambientes de FL [Li et al. 2023, Malandrino and Chiasserini 2021, Hou et al. 2022, Haffar et al. 2022]. Entretanto, estudos sobre como o FL impacta na explicabilidade dos modelos ainda são escassos.

Sob esta motivação, este trabalho propõe uma metodologia para análise da influência do FL na explicabilidade de modelos de IA interpretados com a técnica SHAP. Como objetivo, a metodologia aqui proposta busca explorar as seguintes questões: i) Como o uso de FL impacta a explicabilidade dos modelos? e ii) Quais são os fatores que mais influenciam a explicabilidade? Tais questões são desafiadoras porque o aprendizado federado impõe restrições de acesso aos dados, podendo introduzir heterogeneidade entre clientes, dificultando assim a comparação direta entre modelos federados e centralizados.

Para superar tais desafios, é proposta uma metodologia para comparar as explicações em ambientes centralizados e federados, o que permite evidenciar quando explicações locais e globais são ou não consistentes. O principal resultado desse artigo mostra que o FL altera de forma significativa as explicações, sobretudo em cenários com dados heterogêneos.

O restante deste artigo está organizado da seguinte forma. A Seção 2 faz uma breve revisão do método SHAP. A Seção 3 apresenta trabalhos que exploram o uso integrado de FL e XAI. A Seção 4 descreve a metodologia proposta. Na Seção 5 é apresentada uma aplicação da metodologia usando um *dataset* de referência. Ao final, a Seção 6 faz uma síntese dos resultados e discute extensões futuras do trabalho.

2. IA Explicável com Valores SHAP

SHAP é um dos métodos de XAI mais utilizados para explicar modelos de aprendizado de máquina. Ele se baseia nos valores de Shapley da Teoria dos Jogos [Shapley 1953], e tem como objetivo quantificar de forma justa a contribuição individual de cada característica para uma predição específica do modelo. Uma das principais vantagens do SHAP é sua forte fundamentação matemática, satisfazendo propriedades como simetria e aditividade.

Considerando um modelo de predição f e uma instância $\mathbf{x}$ composta por M características (i.e., $\mathbf{x} \in \mathbb{R}^M$), a predição pode ser decomposta de forma aditiva como

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j$$

onde ϕ_0 representa o valor base da predição, isto é, a saída média esperada do modelo, e ϕ_j corresponde ao valor SHAP associado à instância $\mathbf{x}$ e à característica j, indicando sua contribuição individual para a predição final $f(\mathbf{x})$.

O termo ϕ_0 é geralmente estimado a partir da média das predições do modelo sobre um subconjunto representativo do conjunto de dados. Idealmente, ϕ_0 deve ser calculado sobre todo o *dataset*. Entretanto, devido ao alto custo computacional, é comum o uso de uma amostra suficientemente representativa, o que permite uma boa aproximação do valor base. Já o cálculo dos valores ϕ_j emprega uma perturbação das entradas do modelo. Para a característica j, calcula-se sua contribuição considerando todas as possíveis combinações de subconjuntos de características que não contêm j.

Apesar das vantagens do SHAP, sua aplicação em ambientes de FL apresenta desafios significativos: a dependência do acesso direto aos dados para o cálculo das expectativas marginais colide com a premissa de privacidade do paradigma federado, onde os dados de um dispositivo permanecem locais e inacessíveis ao servidor central e aos demais dispositivos da federação.

3. Trabalhos Relacionados

Os trabalhos que fazem uso de forma conjunta de métodos de XAI e FL geralmente seguem duas linhas: i) uso de XAI para influenciar o comportamento do modelo federado; ou ii) formas de garantir que o modelo federado possa ser explicado. Nesta linha, estão inclusos tanto métodos que envolvem a criação de modelos federados que facilitem sua interpretabilidade quanto a agregação de explicações locais em uma explicação global.

Como exemplo da primeira linha, [Malandrino and Chiasserini 2021] usam XAI para identificar clientes que contribuem negativamente para a acurácia do modelo. Em [Hou et al. 2022], os autores propõem o uso de técnicas de XAI como GradCAM [Selvaraju et al. 2019] para criar filtros de detecção de ataques de *backdoor* por parte de clientes maliciosos. De forma similar, [Haffar et al. 2022] propõem o uso de modelos de XAI para detectar clientes maliciosos em ambientes de FL. Em [Ma and Gu 2023], valores SHAP são utilizados para identificar e mascarar os pixels mais críticos nas imagens de treinamento, garantindo privacidade sem comprometer significativamente a precisão do modelo. No trabalho de [Yuan et al. 2022], os autores usam valores Shapley para melhorar a acurácia em modelos federados de IA para predição de doenças.

Para a segunda linha de pesquisa, em [Wang 2019] os autores propõem o uso de valores Shapley para o cálculo da relevância das características. Estes valores Shapley são determinados localmente e consolidados em um valor único que é enviado ao servidor. Abordagens similares são feitas por [Ducange et al. 2024], com uso de FCM (*Federated Fuzzy Clustering*) para criação do conjunto de dados a ser usado no cálculo do valor base do SHAP; e por [Kalakoti et al. 2025], que agregam valores Shapley usando criptografia diferencial.

Vários autores propõem *frameworks* federados baseados em modelos interpretáveis por definição. Esse é o caso em [Younis et al. 2023, Polato et al. 2022, Imakura et al. 2020, Chen et al. 2021, Dong et al. 2022], que apresentam soluções baseadas em árvores de decisão e criptografia homomórfica para garantir a privacidade dos modelos. O framework EVFL [Chen et al. 2022] combina estratégias de avaliação de credibilidade e explicações contrafactuais federadas [Guidotti 2022]. Já [Liang and Wang 2022] apresenta o FedTSC, um *framework* FL para classificação de séries temporais.

Apesar dos vários trabalhos citados nesta seção, é importante destacar que nenhum deles realiza uma análise do impacto que o FL tem nas explicações e na interpretabilidade dos respectivos modelos. Tal análise é um ponto de partida fundamental na combinação de XAI e FL, e a falta de estudos adequados na literatura é a principal motivação para o desenvolvimento deste trabalho, sendo a metodologia proposta a sua principal contribuição.

4. Metodologia

Esta seção descreve a metodologia proposta para analisar a influência do FL nas explicações geradas por meio de métodos de XAI. Em específico, deseja-se investigar como o FL impacta na estabilidade, coerência e similaridade das explicações baseadas em valores SHAP. Para isso, a estratégia deste trabalho consiste em comparar modelos obtidos por meio de treinamentos centralizados (i.e., não federados) e federados, tomando como referência um mesmo conjunto de dados, e analisando como a interação entre FL e XAI se comporta em diferentes cenários de heterogeneidade de dados.

4.1. Notação e Definições

Tanto para o caso centralizado quanto para o caso federado, a tarefa de aprendizado é a mesma. A entrada para um algoritmo de aprendizado $\mathcal{A}$ é um conjunto de dados de treinamento $D_{\text{Train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, onde $\mathbf{x}_i \in \mathbb{R}^M$ é o vetor de M características (ou *features*) da instância i, e $y_i \in \{0, 1\}$ é o rótulo (ou classe) correspondente. O objetivo é encontrar um classificador f que tenha boa taxa de acerto e bom poder de generalização para dados não vistos durante o treinamento[1]. A diferença principal entre o aprendizado centralizado e o federado consiste na forma de acesso aos dados. No caso centralizado, o algoritmo de treinamento $\mathcal{A}$ tem acesso ao D_{Train} por completo. Já no caso federado, D_{Train} está particionado em N clientes (ou dispositivos). Cada um desses clientes tem acesso somente à sua parte dos dados e ao algoritmo de treinamento. Ao longo do FL, cada cliente conhece o modelo global, obtido (agregado) a partir dos modelos locais de todos os participantes. No entanto, nesse paradigma de treinamento, assume-se que os dados de cada cliente sejam conhecidos apenas por ele mesmo.

Dessa forma, comparar os valores SHAP obtidos no FL com os que seriam obtidos no aprendizado centralizado em situações reais de FL é uma tarefa extremamente desafiadora (ou até mesmo inviável) devido à questão do acesso restrito aos dados. **Por isso, o ponto de partida da metodologia proposta consiste em utilizar um conjunto de dados conhecido.** Com esse conjunto de dados, é possível tanto obter modelos (e computar os valores SHAP) por meio de treinamento centralizado quanto por meio de FL, bastando,

[1]Por simplicidade de notação, esta seção trata apenas da versão binária do problema de classificação. No entanto, todos os conceitos e a metodologia generalizam de forma natural para o caso multiclasse.

nesse último caso, simular o FL. Essa simulação pode ser feita, inclusive, considerando diferentes estratégias de particionamento dos dados, para um dado conjunto de clientes. Com o intuito de formalizar as diferentes formas de aprendizado e as diferentes maneiras de organização dos dados, considere as definições apresentadas a seguir:

- MC, modelo treinado de forma centralizada;
- MF, modelo treinado de forma federada, ou seja, o modelo global após as rodadas de treinamento com dados locais e agregação;
- DT, dados totais, utilizado no treinamento centralizado e na avaliação global (a ser detalhada mais adiante);
- DL, dados locais, no aprendizado federado;
- Uni, cenário, no aprendizado federado, no qual os dados locais de todos os clientes possuem a mesma (ou aproximadamente a mesma) proporção de classes que o conjunto de dados DT;
- $NonUni$, cenário, no aprendizado federado, no qual os clientes têm acesso a dados com proporções de classes significativamente diferentes entre si e também com relação ao conjunto DT.

4.2. Cenários para Análise Comparativa

Foram considerados três cenários base para o cálculo de valores SHAP. O primeiro cenário corresponde ao caso clássico de aprendizado supervisionado centralizado, no qual os dados são apenas divididos em porções para treinamento e teste e o algoritmo de aprendizado tem acesso a todos os dados da porção de treinamento. Formalizando, tem-se: $DT = D_{\text{Train}} \cup D_{\text{Test}}$; $D_{\text{Train}} \cap D_{\text{Test}} = \{\}$; MC é treinado com os dados em D_{Train}; e valores SHAP são computados com dados em D_{Test}. Esse cenário é denotado por MC_DT.

Nos outros dois cenários, relativos ao modelo federado, tem-se que as instâncias correspondentes (ao aprendizado centralizado) são tais que: $D_{\text{Train}} = D_{\text{Train},1} \cup \cdots \cup D_{\text{Train},N}$ e $D_{\text{Test}} = D_{\text{Test},1} \cup \cdots \cup D_{\text{Test},N}$, onde o cliente i tem acesso a apenas $D_{\text{Train},i}$ e $D_{\text{Test},i}$ para os procedimentos de treinamento e teste, respectivamente. Assim, o cenário 2, denotado por MF_DT, é aquele em que o modelo federado resultante, MF, é utilizado para computar os valores SHAP nos elementos de DT_{Test}. Já o cenário 3, denotado por MF_DL, corresponde a uma situação real de FL, onde cada cliente i utiliza o modelo resultante do aprendizado federado, MF, para computar os valores SHAP apenas para a porção dos dados locais de teste a que tem acesso, i.e., para $D_{\text{Test},i}$.

É importante ressaltar que nos três cenários base, os valores SHAP são sempre calculados, de maneiras diferentes, para os mesmos elementos do conjunto de dados, i.e., aqueles em D_{Test}. Esse fato permite a comparação direta dos valores SHAP nos diversos cenários. De posse dos valores SHAP para os cenários descritos anteriormente, são realizadas as análises comparativas descritas a seguir.

1. MF_DT versus MC_DT: tem como objetivo avaliar diretamente a influência do processo de federação (modelo MF) nas explicações, tomando como referência o cenário sem federação (centralizado, modelo MC). Dessa forma, é possível verificar se o treinamento distribuído altera a interpretação do modelo quando os valores SHAP são computados com o mesmo conjunto de dados.
2. MF_DL versus MF_DT: busca investigar se as explicações locais, obtidas a partir dos dados de cada cliente (situação real de FL), diferem das explicações globais que seriam obtidas a partir de MF.

Em ambos os tipos de comparação acima, são considerados os casos em que MF é obtido tanto para o cenário Uni quanto para o $NonUni$.

4.3. Preparação dos Conjuntos de Dados

A premissa principal neste artigo é tal que a metodologia proposta permite a análise da variação dos valores SHAP em diferentes cenários de treinamento e divisão do conjunto de dados. No entanto, é sabido da literatura de aprendizado de máquina que o treinamento de modelos (e por consequência o cômputo dos valores SHAP) pode variar em diferentes divisões de conjuntos de treino e teste e também em diferentes divisões desses conjuntos para os clientes do aprendizado federado. Para contornar o problema de variações nos valores SHAP decorrentes puramente da aleatoriedade das divisões do conjunto de dados, a metodologia deste trabalho propõe repetir as análises em diversas partições de treino e teste. Assim, faz-se possível o cálculo de médias e medidas de dispersão.

As repetições são baseadas no procedimento clássico de validação cruzada. Inicialmente, o conjunto de dados DT é particionado, aleatoriamente, em K *folds*. Para criar um ambiente de experimentação controlado, o qual permita análises focadas apenas no impacto do uso de FL, o processo de divisão considerado é estratificado, ou seja, as distribuições de classes em todos os *folds* são as mesmas (ou muito similares).

Na rodada k ($k = 1, \ldots, K$) da validação cruzada, o k-ésimo *fold* é utilizado para teste e cômputo dos valores SHAP e os demais $K - 1$ *folds* são utilizados para os treinamentos dos modelos. Com isso, os conjuntos D_{Train} e D_{Test} (descritos na Seção 4.2) para a rodada k são denotados por D^k_{Train} e D^k_{Test}.

Para cada k, o conjunto de treinamento D^k_{Train} é subdividido em N subconjuntos, onde N representa o número de clientes participantes do aprendizado federado. Da mesma forma, o conjunto de teste D^k_{Test} também é subdividido entre os mesmos N clientes. Logo:

$$D^k_{\text{Train}} = \bigcup_{i=1}^{N} D^k_{\text{Train},i} \qquad \text{e} \qquad D^k_{\text{Test}} = \bigcup_{i=1}^{N} D^k_{\text{Test},i},$$

onde $D^k_{\text{Train},i}$ e $D^k_{\text{Test},i}$ representam, respectivamente, os subconjuntos de treino e teste atribuídos ao cliente i durante a rodada k da validação cruzada.

Para as subdivisões dos conjuntos D^k_{Train} e D^k_{Test} entre os N clientes, propõe-se: no cenário Uni, particionar tais conjuntos de maneira aleatória em N partes de mesmo tamanho e com as mesmas (ou praticamente as mesmas) distribuições de classes; no cenário *NonUni*, os tamanhos dos N subconjuntos são definidos a partir de uma distribuição de Dirichlet, gerando clientes com visões variadas do conjunto de dados (tanto no tamanho dos subconjuntos de treinamento e teste, quanto nas distribuições de classes observadas). No entanto, no caso *NonUni*, também é necessário garantir a viabilidade do treinamento. Para isso, são adicionadas duas restrições: i) cada cliente deve possuir um número mínimo de amostras; e ii) não são permitidos clientes que contenham dados de uma única classe.

4.4. Treinamento dos Modelos e Cálculo dos Valores SHAP

Para a k-ésima rodada da validação cruzada ($k = 1, \ldots, K$), três modelos são treinados: um modelo centralizado; um modelo federado no cenário *Uni*; e um modelo federado no cenário *NonUni*. Durante o treinamento de cada modelo, são registradas métricas de

desempenho relevantes para a tarefa de classificação, como acurácia e F1-*score*. Essas métricas são armazenadas individualmente para os respectivos *folds* de teste e são calculados a média e o desvio padrão delas ao longo dos K *folds*. Além do treinamento e da avaliação, para cada rodada da validação cruzada, os valores SHAP e os valores-base SHAP são calculados de acordo com os cinco cenários apresentados na Seção 4.2: MC_DT, MF_DT (*Uni* e *NonUni*) e MF_DL (*Uni* e *NonUni*).

4.5. Medidas de Comparação

Para quantificar as diferenças entre os valores SHAP nos diversos cenários apresentados nas últimas seções, este trabalho propõe o uso de algumas funções de distância e similaridade. Antes de apresentar tais funções, considere as seguintes definições:

- $\boldsymbol{\varphi}_C^k(\mathbf{x})$, o vetor de valores SHAP (para as M características do problema de classificação) obtido na k-ésima rodada da validação cruzada, para o cenário C (conforme Seções 4.2 e 4.4) e para a instância $\mathbf{x} \in D_{\text{Test}}^k$;
- $\mathbb{I}_{C,\ell}^k(\mathbf{x})$, o conjunto dos índices dos ℓ elementos com maiores valores absolutos no vetor $\boldsymbol{\varphi}_C^k(\mathbf{x})$. Esse conjunto representa as ℓ características mais relevantes para a classificação do elemento $\mathbf{x}$, conforme o modelo utilizado.

Assim, propõe-se comparar os valores SHAP das seguintes maneiras, considerando k, $\mathbf{x}$ e ℓ fixos e cenários distintos C e C':

1. Calcular a **Distância do Cosseno** entre $\boldsymbol{\varphi}_C^k(\mathbf{x})$ e $\boldsymbol{\varphi}_{C'}^k(\mathbf{x})$. A distância do cosseno varia entre 0 e 2, e quanto maior é o valor da distância, maior é o ângulo entre os dois vetores. A ideia de se utilizar essa distância é quantificar o quão diferentes são as distribuições de valores SHAP nas M características nos cenários C e C'.
2. Calcular a **Coincidência de Características** entre $\mathbb{I}_{C,\ell}^k(\mathbf{x})$ e $\mathbb{I}_{C',\ell}^k(\mathbf{x})$, representada pelo número de elementos da interseção entre esses conjuntos. O valor de ℓ escolhido pode depender do número total de características, M. No entanto, sugere-se utilizar valores pequenos de ℓ (e.g., 5, 10 e 15) para comparar apenas as top-ℓ características nos dois cenários.
3. Calcular a **Consistência do Sinal** das características mais relevantes nos cenários C e C'. Para isso, propõe-se calcular a fração das características em $\mathbb{I}_{C,\ell}^k(\mathbf{x}) \cap \mathbb{I}_{C',\ell}^k(\mathbf{x})$ que possuam o mesmo sinal em $\boldsymbol{\varphi}_C^k(\mathbf{x})$ e $\boldsymbol{\varphi}_{C'}^k(\mathbf{x})$. A intuição é verificar o quão comumente as características mais relevantes contribuem para que as classes preditas de $\mathbf{x}$ sejam as mesmas nos cenários C e C'.

5. Estudo de Caso: Aplicação da Metodologia no *Dataset* EHMS

A metodologia foi aplicada usando o conjunto de dados *Edge Health Monitoring System* (EHMS)[2] [Hady et al. 2020], voltado para detecção de ataques em sistemas de monitoramento de saúde, combinando dados diversos: características de tráfego de rede e dados de sinais biométricos. Em comparação a outras bases de dados disponíveis na literatura, o EHMS apresenta menor escala e boa diversidade de tipos de dados, o que possibilitou a validação da metodologia proposta de forma mais controlada e com menor custo computacional.

[2] `https://www.cse.wustl.edu/~jain/ehms/index.html`

Tabela 1. Métricas de desempenho dos modelos.

	Média		Desvio Padrão	
Cenário	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**
MC	0.9094	0.7832	0.0092	0.0218
$MF\ (Uni)$	0.9126	0.7767	0.0130	0.0086
$MF\ (NonUni)$	0.9184	0.7815	0.0079	0.0083

Tabela 2. Valores base (ϕ_0) por *fold* para os cenários em DT.

Cenário	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Média**	**DP**
MC_DT	-1.124	-1.334	-0.085	0.023	-0.705	-0.645	0.606
$MF\_DT\ (Uni)$	-0.046	-0.036	-1.082	0.154	-0.422	-0.286	0.492
$MF\_DT\ (NonUni)$	-5.488	-0.472	-1.834	-2.029	-1.384	-2.241	1.912

5.1. Treinamento dos Modelos

O treinamento do modelo centralizado (MC) foi realizado utilizando *PyTorch*[3]. Já os treinamentos federados foram feitos por meio do framework *MininetFed*[4] [Sarmento et al. 2024], permitindo a emulação controlada de cenários de aprendizado federado com múltiplos clientes. Todos os experimentos de federação foram realizados com $N = 4$ clientes. Tal valor foi escolhido por ser considerado um número mínimo de clientes representativo para a simulação adequada de um ambiente federado, e também por permitir a apresentação dos resultados de forma inteligível.

A arquitetura utilizada no modelo de classificação é a mesma proposta originalmente em [Hady et al. 2020]. Para fins de rotação de conjuntos de treinamento e teste, foi utilizada a técnica de validação cruzada com $K = 5$ *folds* (vide Seção 4.3). As métricas de desempenho dos modelos treinados (Média e Desvio Padrão – DP, relativos aos 5 conjuntos de teste) são apresentadas na Tabela 1. É importante observar que não é o objetivo aqui garantir a obtenção do modelo mais eficaz. Ainda assim, os resultados para o treinamento centralizado são compatíveis com os alcançados por [Hady et al. 2020], o que serve como validação para a etapa de treino. Em todos os cenários, os desvios padrão entre os *folds* foram baixos, indicando estabilidade do treinamento.

5.2. Cálculo dos Valores SHAP

Após a etapa de treinamento, os valores SHAP foram calculados para os cenários descritos na Seção 4.2, em todas as instâncias de cada *fold* de teste. A Tabela 2 apresenta os valores base do SHAP (ϕ_0) para os cenários MC_DT e MF_DT (Uni e $NonUni$). Pode-se observar uma grande variação desses valores entre os *folds*, em todos os cenários. Isso evidencia: i) que os modelos resultantes são diferentes entre si, mesmo quando estes alcançam métricas de desempenho semelhantes; e ii) que a distribuição dos dados entre os *folds* influencia na formação de ϕ_0.

[3] `https://pytorch.org/`

[4] `https://github.com/lprm-ufes/MininetFed`

Tabela 3. Valores base (ϕ_0) por *fold* e por cliente para $MF\_DL\ (Uni)$.

Cliente	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Média	DP
1	-0.282	-0.140	-1.236	0.866	-0.385	-0.235	0.750
2	0.689	-0.140	-1.026	-0.206	-0.201	-0.177	0.607
3	-0.072	-0.066	-1.272	0.217	-0.352	-0.309	0.575
4	-0.133	0.531	-1.151	-0.093	-0.001	-0.169	0.610

Tabela 4. Valores base (ϕ_0) por *fold* e por cliente para $MF\_DL\ (NonUni)$.

Cliente	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Média	DP
1	-7.151	-0.998	-2.944	-2.430	-2.067	-3.118	2.364
2	-0.889	0.921	-0.416	-0.396	0.867	0.017	0.825
3	-3.878	-0.366	-1.643	-2.292	-1.295	-1.895	1.309
4	-4.826	-1.113	-2.847	-2.789	-2.015	-2.718	1.372

As Tabelas 3 e 4 apresentam, para cada *fold* e cliente federado, os valores de ϕ_0 para os modelos usando somente os dados locais, i.e., MF_DL (Uni e $NonUni$). Novamente observa-se grande variação de ϕ_0. Esses resultados sugerem que os modelos diferem entre si, seja por conta da distribuição dos dados locais entre clientes ou por conta de como essa distribuição foi realizada (Uni e $NonUni$).

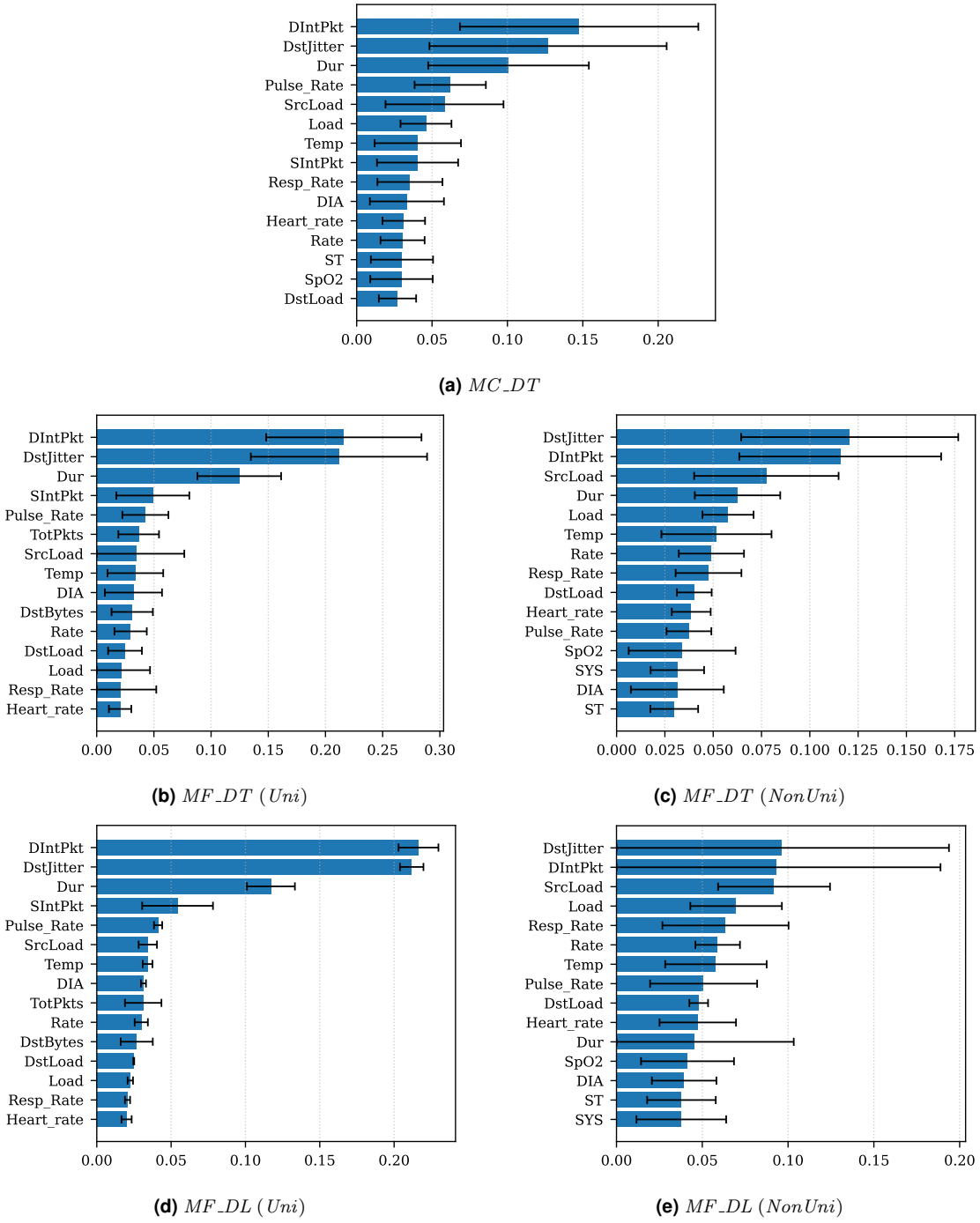
A Figura 1 mostra as explicações globais[5] SHAP para as 15 características mais importantes. A variação dos intervalos de confiança é alta, decorrente da variabilidade entre os *folds*. Ao comparar os cenários, observa-se que existem diferenças significativas nas explicações globais: embora algumas características permaneçam entre as mais importantes, a ordem relativa varia entre os cenários. Além disso, no cenário Uni a importância tende a se concentrar em um subconjunto menor de características, enquanto no cenário $NonUni$ a importância se distribui de forma mais homogênea.

5.3. Comparação de Cenários

Foram calculadas as distâncias de cosseno, as coincidências de características e a consistência do sinal das explicações, para fins de comparação entre as explicações dos cenários definidos na Seção 4.2. Na apresentação dos resultados foram utilizados gráficos para curva de função de distribuição cumulativa empírica (ECDF – *Empirical Cumulative Distribution Function*) para verificar a similaridade entre as distâncias de cossenos, e gráficos usando função de massa de probabilidade (PMF – *Probability Mass Function*) para as coincidências de características e consistência de sinais.

A Figura 2 mostra as curvas ECDF, por *fold*, na comparação das explicações usando a distância de cosseno. Na comparação MF_DT vs. MC_DT, apenas no *fold* 4 do cenário Uni o modelo federado apresenta comportamento mais próximo ao centralizado. No cenário $NonUni$, nenhum *fold* se destacou em alta similaridade. Esses resultados, em conjunto com as análises do valor ϕ_0, reforçam que a distribuição dos dados exerce forte influência sobre as explicações SHAP e que o aprendizado federado produziu modelos distintos dos centralizados, mesmo com métricas de desempenho semelhantes.

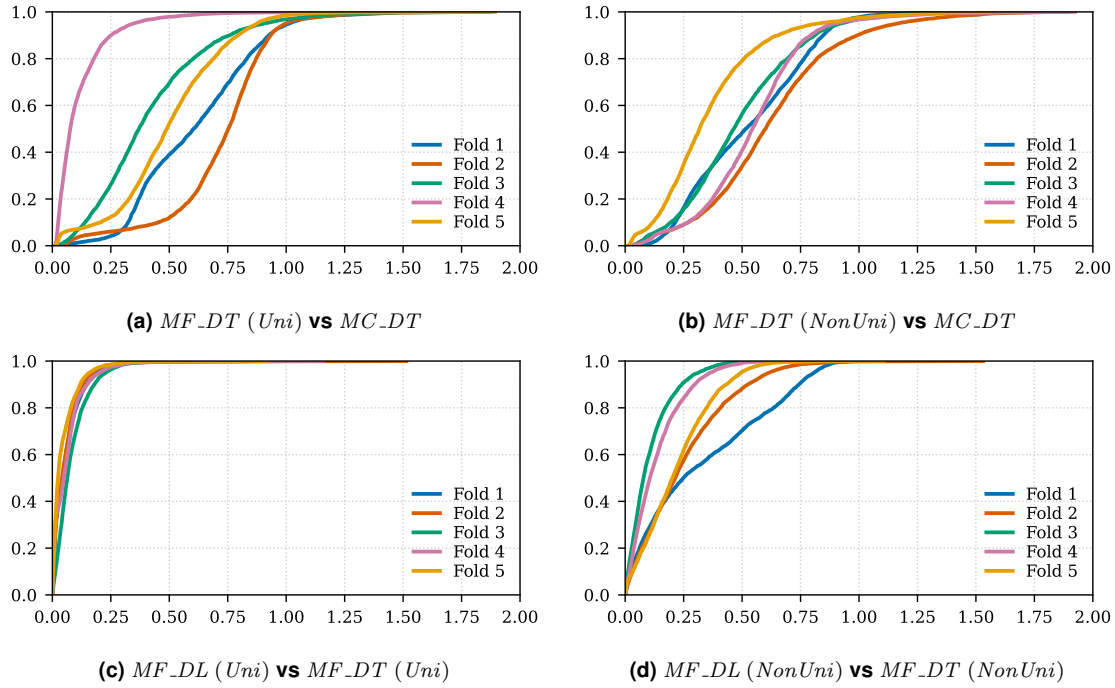
[5]Média dos valores absolutos dos valores SHAP para todas as instâncias dos *folds* de teste.



(a) MC_DT **(b)** $MF\_DT\ (Uni)$ **(c)** $MF\_DT\ (NonUni)$ **(d)** $MF\_DL\ (Uni)$ **(e)** $MF\_DL\ (NonUni)$

Figura 1. Top-15 características segundo a importância global (médias dos valores absolutos) dos valores SHAP. Linhas horizontais representam Intervalos de Confiança de 95%.

Na comparação $MF\_DL\ (Uni)$ vs. $MF\_DT\ (Uni)$, observa-se que as explicações são altamente similares em todos os *folds*. Como a distribuição dos dados locais replica a distribuição global, cada cliente é capaz de explicar o modelo de forma consistente utilizando apenas seus dados. **Nesse caso, não há, por exemplo, necessidade de técnicas para agregação de explicações entre clientes.** Esse comportamento, entretanto, não se



(a) $MF\_DT\ (Uni)$ **vs** MC_DT **(b)** $MF\_DT\ (NonUni)$ **vs** MC_DT

(c) $MF\_DL\ (Uni)$ **vs** $MF\_DT\ (Uni)$ **(d)** $MF\_DL\ (NonUni)$ **vs** $MF\_DT\ (NonUni)$

Figura 2. Curvas ECDF para a distância do cosseno entre explicações SHAP.

mantém no cenário $NonUni$, onde a distribuição heterogênea dos dados impacta diretamente as explicações locais.

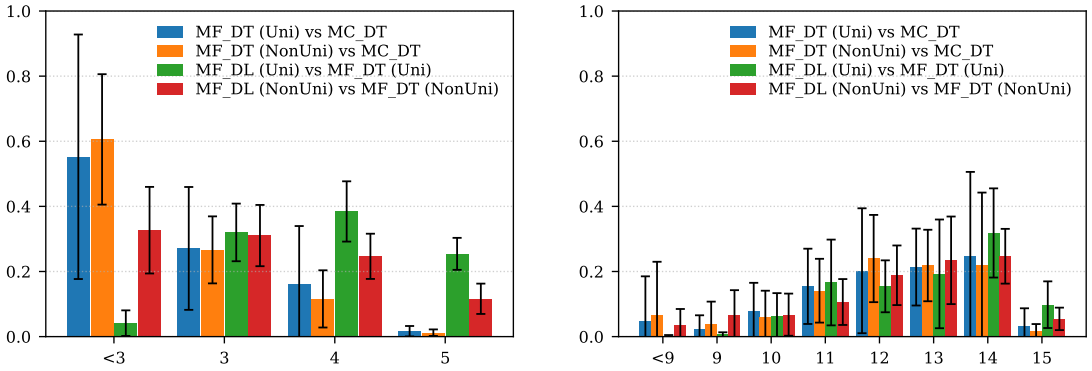
De forma geral, nota-se que a similaridade entre explicações é significativamente maior nos cenários que comparam modelos federados entre si (MF_DL vs. MF_DT Uni/$NonUni$) do que nos cenários que usam como base os modelos centralizados (MF_DT vs. MC_DT Uni/$NonUni$). Isso sugere que a federação, por si só, produz modelos distintos, os quais naturalmente produzem também explicações diferentes.

A similaridade de rankings também mostra essas diferenças. A Figura 3 apresenta a PMF da quantidade de características com rankings coincidentes para as 5 e 15 características mais importantes. Para as 5 principais características, o cenário MF_DT vs. MC_DT apresenta maior prevalência de apenas duas posições coincidentes, enquanto MF_DL vs. MF_DT apresenta maior concordância. Esse comportamento se mantém para 15 características, embora a inclusão de características menos relevantes reduza a diferença.

Por fim, a Figura 4 mostra a PMF da quantidade de sinais consistentes entre os rankings. Novamente, as explicações produzidas por MF_DL e MF_DT são mais similares e apontam para a mesma direção. Na comparação entre modelos federados e centralizados (MF_DT vs. MC_DT), há alta consistência para as 5 características mais importantes, mas isso diminui à medida que mais características são consideradas.

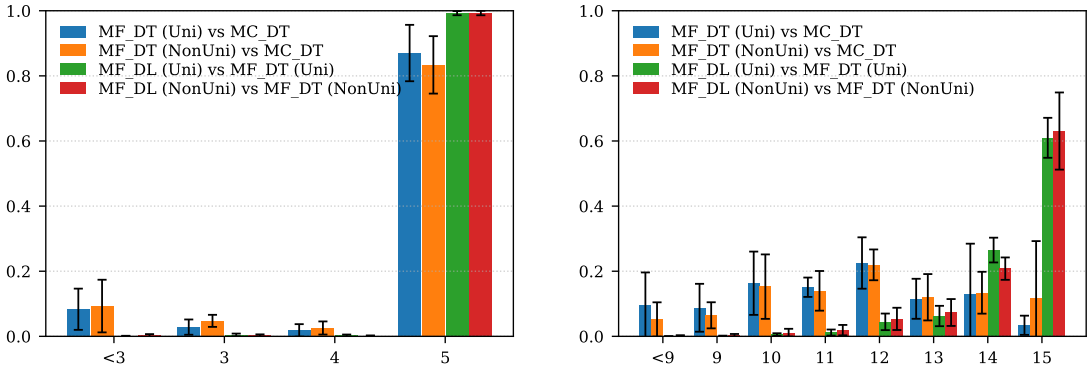
5.4. Sumário dos Resultados Experimentais

Considerando os resultados obtidos, observa-se que o ambiente federado exerce influência significativa sobre as explicações SHAP, uma vez que o modelo resultante da federação difere daquele obtido por treinamento centralizado, o que naturalmente leva a explicações distintas. Além disso, a distribuição dos dados entre os clientes impacta de forma re-



(a) Top-5 **(b) Top-15**

Figura 3. Distribuição de probabilidade (PMF) do número de características coincidentes entre os rankings SHAP no Top-ℓ, com intervalo de confiança de 95%.



(a) Top-5 **(b) Top-15**

Figura 4. Distribuição de probabilidade (PMF) da quantidade de sinais consistentes entre os rankings SHAP no Top-ℓ, com intervalo de confiança de 95%.

levante a consistência das explicações. Em ambientes federados com distribuição de dados homogênea (Uni), as explicações locais mostram-se equivalentes às explicações globais, reduzindo a necessidade de técnicas complexas de agregação. Por outro lado, em cenários heterogêneos ($NonUni$), as explicações locais tornam-se divergentes, o que justifica a adoção de estratégias de agregação entre clientes. Por fim, caso o objetivo seja obter explicações federadas mais próximas das centralizadas, torna-se necessário empregar técnicas que aproximem o modelo federado daquele treinado de forma centralizada.

6. Conclusões e Trabalhos Futuros

Neste trabalho foi proposta uma metodologia para análise da influência de ambientes de aprendizado federado sobre explicações baseadas em valores SHAP.

A abordagem foi validada por meio de sua aplicação ao conjunto de dados de referência EHMS. Os resultados obtidos demonstram que o ambiente federado exerce forte influência sobre as explicações via valores SHAP, mesmo em cenários nos quais os modelos federados e centralizados apresentam métricas de desempenho semelhantes. Além disso, foi identificado que a distribuição dos dados entre os clientes também tem impacto significativo nas explicações. Quando a distribuição local dos dados em cada cliente é

similar à distribuição global do conjunto de dados, não há diferença significativa entre explicações locais e globais. Dessa forma, não são necessárias técnicas de agregação de explicações entre clientes. Por outro lado, quando a distribuição dos dados locais difere da distribuição global (situação mais próxima da realidade), as explicações locais e globais deixam de apresentar correspondência, evidenciando a necessidade de mecanismos específicos para lidar com essa diferença.

Como trabalhos futuros, pretende-se aplicar a metodologia proposta a outros conjuntos de dados, a fim de verificar se os achados observados para o EHMS se mantêm em outros conjuntos de dados. Além disso, planeja-se estender a metodologia para realizar uma análise temporal das explicações, investigando como o ambiente federado impacta os valores SHAP ao longo dos *rounds* de treinamento. Por fim, destaca-se como direção futura o estudo de estratégias para agregação de explicações entre clientes, buscando garantir que explicações locais sejam compatíveis com explicações globais, sem violar os requisitos de privacidade de dados característicos de ambientes de aprendizado federado.

Agradecimentos

Este trabalho possui financiamento parcial das seguintes agências: Fapes (2026-B74MN, 2023-RWXSZ, 2025-1H3FP); MCTI/Fapesp/CGI.br (Porvir-5G); CNPq; e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Bak, M. et al. (2024). Federated learning is not a cure-all for data ethics. *Nature Machine Intelligence*, 6.

Branson, J. et al. (2020). Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations. *Trials*, 21.

Chen, P. et al. (2022). EVFL: An explainable vertical federated learning for data-oriented artificial intelligence systems. *Journal of Systems Architecture*, 126:102474.

Chen, X. et al. (2021). Fed-EINI: An efficient and interpretable inference framework for decision tree ensembles in federated learning.

Dong, T. et al. (2022). An interpretable federated learning-based network intrusion detection framework.

Ducange, P. et al. (2024). Consistent post-hoc explainability in federated learning through federated fuzzy clustering. In *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–10.

Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38:1–55.

Hady, A. A. et al. (2020). Intrusion detection system for healthcare systems using medical and network data: A comparison study. *IEEE Access*, 8:106576–106584.

Haffar, R. et al. (2022). Explaining predictions and attacks in federated learning via random forests. *Applied Intelligence*, 53:1–17.

Hou, B. et al. (2022). Mitigating the backdoor attack by federated filters for industrial IoT applications. *IEEE Transactions on Industrial Informatics*, 18(5):3562–3571.

Imakura, A. et al. (2020). Interpretable collaborative data analysis on distributed data.

Kalakoti, R. et al. (2025). Federated learning of explainable AI(FedXAI) for deep learning-based intrusion detection in IoT networks. *Computer Networks*, 270:111479.

Li, A. et al. (2023). Towards interpretable federated learning.

Liang, Z. and Wang, H. (2022). FedTSC: a secure federated learning system for interpretable time series classification. *Proc. VLDB Endow.*, 15(12):3686–3689.

Linardatos, P. et al. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1).

Lopez-Ramos, L. M. et al. (2024). Interplay between federated learning and explainable artificial intelligence: a scoping review.

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

Ma, X. and Gu, L. (2023). Research and application of generative-adversarial-network attacks defense method based on federated learning. *Electronics*, 12:975.

Malandrino, F. and Chiasserini, C. F. (2021). Toward node liability in federated learning: Computational cost and network overhead. *IEEE Communications Magazine*, 59(9):72–77.

Markus, A. F. et al. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655.

McMahan, H. B. et al. (2023). Communication-efficient learning of deep networks from decentralized data.

Polato, M. et al. (2022). Boosting the federation: Cross-silo federated learning without gradient descent. In *2022 Int Joint Conf on Neural Networks (IJCNN)*, pages 1–10.

Sarmento, E. M. et al. (2024). MininetFed: A tool for assessing client selection, aggregation, and security in federated learning. In *2024 IEEE 10th World Forum on Internet of Things (WF-IoT)*, pages 1–6. IEEE.

Selvaraju, R. R. et al. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Shapley, L. S. (1953). *A Value for n-Person Games*, pages 307–318. Princeton Univ. Press.

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

Yang, Q. et al., editors (2020). *Federated Learning - Privacy and Incentive*, volume 12500 of *Lecture Notes in Computer Science*. Springer.

Younis, R. et al. (2023). FLAMES2Graph: An interpretable federated multivariate time series classification framework. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 3140–3150. ACM.

Yuan, X. et al. (2022). An efficient digital twin assisted clustered federated learning algorithm for disease prediction. In *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, pages 1–6.