UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

GUILHERME CAROLINO RODRIGUES E ROCHA

COMPATIBILIDADE ENTRE CLIENTES PARA PREDIÇÃO DE DESEMPENHO DO APRENDIZADO FEDERADO

Ouro Preto
2026

GUILHERME CAROLINO RODRIGUES E ROCHA

COMPATIBILIDADE ENTRE CLIENTES PARA PREDIÇÃO DE DESEMPENHO DO APRENDIZADO FEDERADO

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Dr. Rodrigo César Pedrosa Silva
Coorientador: B.Sc. Vinícius Nascimento Targa

Ouro Preto
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Guilherme Carolino Rodrigues e Rocha

Compatibilidade entre Clientes para Predição de Desempenho do Aprendizado Federado

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 26 de Fevereiro de 2026

Membros da banca

Rodrigo César Pedrosa Silva (Orientador) - Doutor - Universidade Federal de Ouro Preto
Vinícius Nascimento Targa (Coorientador) - Bacharel - Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Ouro Preto
Daniel Ludovico Guidoni (Examinador) - Doutor - Universidade Federal de Ouro Preto
Ray da Silva Basilio (Examinador) - Bacharel - Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Ouro Preto

Rodrigo César Pedrosa Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 26/02/2026

Documento assinado eletronicamente por **Rodrigo Cesar Pedrosa Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/03/2026, às 11:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1062465** e o código CRC **DFF5FA60**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Agradecimentos

Agradeço ao meu orientador, Dr. Rodrigo César Pedrosa Silva, e ao meu coorientador, B.Sc. Vinícius Nascimento Targa, pela dedicação e pelo suporte constante que possibilitaram a realização deste trabalho.

Agradeço também aos meus amigos e familiares que me acompanharam durante toda essa caminhada. Em especial, à minha namorada Isadora, ao meu pai Sebastião Márcio, à minha mãe Márcia Junia Carolino, aos meus irmãos Pedro e Gustavo, à minha tia Di e meu tio José Nilton, ao meu tio Marcelo, à minha tia Cristina, à minha avó Agenília, ao meu tio José Márcio, ao meu tio Gutinho e ao meu primo Davi.

Aos amigos da Batcaverna, em especial ao Leonardo, que compartilhou inúmeros momentos do curso comigo.

Resumo

O aprendizado federado permite o treinamento colaborativo de modelos sem o compartilhamento direto de dados, porém enfrenta desafios significativos decorrentes da heterogeneidade entre clientes. Neste trabalho, introduzimos o conceito de compatibilidade entre clientes, definido como a discrepância estatística entre o modelo global de uma rodada anterior e os modelos locais atualizados, avaliada sobre os dados de cada cliente nas rodadas iniciais do treinamento. A compatibilidade é estimada por meio de métricas de divergência e distância entre distribuições, incluindo Maximum Mean Discrepancy (MMD), Fréchet Inception Distance (FID), divergência de Kullback-Leibler (KL) e divergência de Jensen-Shannon (JS). Investigamos a correlação entre a compatibilidade calculada nas primeiras rodadas e o ganho global de acurácia ao final do treinamento federado. Os experimentos foram conduzidos em diferentes conjuntos de dados (CIFAR-10, Fashion-MNIST e Blood-MNIST), variando número de clientes, arquitetura de rede neural e número de épocas locais. Os resultados indicam que a MMD apresentou comportamento consistente e robusto em todos os cenários analisados, mantendo forte correlação negativa com o ganho global. Em contraste, a FID demonstrou instabilidade em determinados contextos, enquanto KL e JS apresentaram comportamento dependente do conjunto de dados e do critério de agregação adotado. Além da análise retrospectiva, demonstramos o potencial uso preditivo da compatibilidade baseada em MMD como indicador antecipado do desempenho final da federação, possibilitando aplicações em diagnóstico precoce e ajuste dinâmico de hiperparâmetros. Os resultados sugerem que a compatibilidade entre clientes, especialmente quando mensurada via MMD, constitui ferramenta promissora para análise e monitoramento de cenários federados heterogêneos.

Palavras-chave: Aprendizado Federado. Transferabilidade. Non-IID .

Abstract

Federated learning enables collaborative model training without direct data sharing, but faces significant challenges due to client heterogeneity. In this work, we introduce the concept of client compatibility, defined as the statistical discrepancy between the global model from a previous round and locally updated client models, evaluated on each client's data during the initial training rounds. Compatibility is estimated using distributional divergence and distance metrics, including Maximum Mean Discrepancy (MMD), Fréchet Inception Distance (FID), Kullback-Leibler (KL) divergence, and Jensen-Shannon (JS) divergence. We analyze the correlation between early-round compatibility and the final global accuracy gain achieved by the federated model. Experiments were conducted across multiple datasets (CIFAR-10, Fashion-MNIST, and Blood-MNIST), varying the number of clients, neural network architectures, and local training epochs. Results indicate that MMD consistently exhibited robust negative correlation with final performance across all scenarios. In contrast, FID showed instability in specific contexts, while KL and JS presented dataset- and aggregation-dependent behavior. Beyond retrospective analysis, we demonstrate the predictive potential of MMD-based compatibility as an early indicator of federated performance, enabling applications such as early diagnosis of unstable federations and dynamic hyperparameter adjustment. Overall, the findings suggest that client compatibility—particularly when measured via MMD—constitutes a promising tool for analyzing and monitoring heterogeneous federated learning systems.

Keywords: Federated Learning, Transferability, Non-IID.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

1 Introdução

Uma quantidade massiva de dados é gerada diariamente em diversos dispositivos eletrônicos como: celulares, *tablets*, *smart-watches*, *wearables*, etc . Utilizar esta grande quantidade de dados para treinar modelos de aprendizado de máquina possibilita a produção de modelos muito poderosos, os quais podem causar impactos significativos na tecnologia e por consequência na sociedade (POUSHTER et al., 2016).

Entretanto, (BAO; GUO, 2021) mostra que em decorrência de preocupações com segurança e privacidade, diversas leis de proteção de dados foram instituídas. Como consequência, o aprendizado de máquina centralizado tem enfrentado diversos desafios, pois muitas vezes não é possível ou não é permitido coletar dados de diversas fontes. Desta forma, mesmo com grande volume de dados existentes não é possível utilizá-los.

Um exemplo da restrição enfrentada pelo aprendizado centralizado é o seguinte: deseja-se criar um modelo de aprendizado de máquina que auxilie a escrita de processos jurídicos. Para isso é necessário coletar diversos processos e fornecê-los como entrada para o treinamento do modelo, entretanto muitos usuários mesmo que desejem o auxílio de escrita, não estão dispostos ou não podem compartilhar o conteúdo de seus processos com terceiros, desta forma o projeto é inviabilizado.

O aprendizado federado do inglês *federated learning* (FL) é uma abordagem de aprendizado de máquina colaborativo na qual o próprio dispositivo gerador de dados realiza computações de treinamento no modelo. Nessa arquitetura, um servidor coordenador hospeda um modelo de aprendizado de máquina global, que é distribuído aos participantes da federação, conhecidos como clientes (MCMAHAN et al., 2017).

O modelo global é um ponto de partida, o servidor inicia o modelo e o distribui aos clientes. O treinamento ocorre de forma local, utilizando os dados presentes nos dispositivos dos clientes. Ao final dessa etapa, um novo modelo local é gerado, e apenas seus parâmetros são transmitidos de volta ao servidor coordenador.

No servidor, os parâmetros recebidos dos diferentes clientes são agregados para gerar uma nova versão do modelo global. Esta nova versão é, então, distribuída novamente aos clientes. Esse processo é definido como uma rodada. A principal vantagem dessa abordagem é a garantida de privacidade dos dados locais dos clientes, uma vez que os dados brutos nunca deixam os dispositivos dos usuários, apenas os parâmetros do modelo local são compartilhados.

Além disso pode ser uma facilidade não precisar coletar os dados todos em um local, evitando a necessidade de um servidor robusto para armazenar e processar estes dados. Isso difere do aprendizado de máquina centralizado tradicional, no qual os dados de todos os clientes

precisam ser enviados e armazenados em um servidor central para o treinamento (BANABILAH et al., 2023). Fato que comprova a relevância do aprendizado federado, pois ele pode viabilizar projetos que não aconteceriam na abordagem centralizada.

Um caso de grande sucesso de aplicação do aprendizado federado é o *Gboard*, um teclado virtual desenvolvido pela *Google*, que disponibiliza algumas funcionalidades como: previsão de texto, correção automática, reconhecimento de voz, etc. Como apresentado por (MCMAHAN; THAKURTA, 2022) a *Google* utilizou aprendizado federado para melhorar diversas funcionalidades do *Gboard*.

Entretanto, é necessário destacar que existem riscos associados à participação em um esquema de aprendizado federado. Quando os clientes da federação comunicam as atualizações de parâmetros de seus modelos ao servidor central, a privacidade pode ser comprometida (RODRíGUEZ-BARROSOA et al., 2022). Além disso, agentes maliciosos podem explorar esse processo para violar a privacidade dos participantes ou inserir dados contaminados, comprometendo o desempenho e a confiabilidade do modelo global (LYU; YU; YANG, 2020). Desta forma, avaliar previamente a viabilidade de participar de uma federação é fundamental para garantir que os benefícios esperados superem os riscos envolvidos.

Ainda que o aprendizado federado tenha possibilitado diversos casos de sucesso, um desafio fundamental ameaça a viabilidade de determinadas federações. A natureza *Non-IID* (não independentemente e identicamente distribuídos) dos dados entre os participantes pode levar a um modelo global com desempenho insatisfatório (ZHAO et al., 2018). Nesse cenário, um cliente pode assumir os riscos inerentes à colaboração federada e, ainda assim, não obter resultados que atendam às suas expectativas.

Dessa forma, torna-se de extrema valia dispor de métodos capazes de indicar a um cliente c_i a viabilidade de sua participação em uma federação F. Uma avaliação prévia desse tipo possibilita uma tomada de decisão mais segura, mitigando riscos e evitando custos computacionais e de privacidade desnecessários.

Com esse objetivo, este trabalho introduz o conceito de compatibilidade entre clientes, definido como o grau de similaridade entre as distribuições de dados dos participantes de uma federação. Essa compatibilidade é aferida por meio de métricas de transferabilidade e distância entre distribuições. A hipótese é que clientes considerados mais compatíveis, apresentem distribuições de dados mais semelhantes e desta forma tendem a compor federações que apresentam melhor desempenho global na tarefa de aprendizado.

1.1 Trabalhos relacionados

No trabalho (ZHU et al., 2021) foi proposto um esquema de aprendizado federado assíncrono, estruturado em duas etapas e com taxa de aprendizado adaptativa, com o objetivo de

mitigar os impactos causados por dados *Non-IID* e heterogeneidade na capacidade de computação e comunicação dos clientes.

Os autores focam em dois desafios principais: a distribuição de dados *Non-IID* e o fenômeno de *staleness*, que representa a defasagem entre os gradientes locais e o estado atual do modelo global. É proposto o algoritmo *WKAFL*, que seleciona gradientes consistentes, mesmo com certo grau de defasagem, e ajusta a taxa de aprendizado de acordo com esse nível. Testes com o conjunto *EMNIST*, sob diferentes graus de *Non-IID* e defasagem, demonstraram ganhos em velocidade de treinamento e precisão.

No entanto, o algoritmo apresenta instabilidade quando os níveis de defasagem dos gradientes são muito elevados ou os dados extremamente heterogêneos.

O trabalho de (SU; XU; YANG, 2023) analisa os algoritmos FedAvg e FedProx sob uma perspectiva estatística, demonstrando que, mesmo na ausência de convergência para pontos estacionários, os modelos aprendidos podem alcançar taxas estatisticamente ótimas. Além disso, o estudo introduz o conceito de *federation gain*, no qual são caracterizadas as condições estatísticas sob as quais um cliente se beneficia da participação em uma federação. Entretanto, apesar dessa caracterização teórica, o trabalho não propõe um mecanismo operacional que permita estimar, nos estágios iniciais do treinamento, o desempenho final de uma federação específica.

(FAMá et al., 2021) analisaram técnicas de seleção de clientes baseadas em métricas de similaridade para cenários com clientes heterogêneos e dados *Non-IID* , com o objetivo de aumentar a velocidade de convergência do modelo e desta forma diminuir o consumo de energia dos dispositivos dos clientes participantes de uma federação.

Assume-se que a distribuição de rótulos é conhecida em cada cliente, é calculado então o número de amostras por rótulo para cada cliente, e divide-se pelo total de amostras. Desta forma é computada a distribuição de probabilidade local. Realizando este cálculo para todos os clientes é obtida a distribuição de probabilidade global de classes.

É proposta então uma estratégia de agrupamento de clientes utilizando métricas de similaridade. A formação dos agrupamentos é baseada nas seguintes métricas de similaridade : função cosseno, erro quadrático médio, distâncias euclidiana, Manhattan,Chebyshev, Wasserstein, discrepância média máxima, divergências de Kullback-Leibler e Jensen-Shannon.

Estas métricas possibilitam o encontro de correlações nas informações locais para formar agrupamentos semânticos. Ao selecionar clientes desses clusters a informação redundante é diminuida, melhorando assim a eficiência da federação.

Por linhas gerais, embora estes trabalhos explorem estratégias eficientes para reduzir a problemática dos efeitos de distribuições heterogêneas, seja por meio de algoritmos assíncronos, análises estatísticas ou seleção de clientes baseada em similaridade, observa-se que a literatura carece de mecanismos práticos capazes de estimar, nos estágios iniciais do treinamento, a viabilidade de uma federação.

1.2 Objetivos

1.2.1 Objetivo geral

Analisar, de forma empírica, se a compatibilidade entre clientes em um esquema de aprendizado federado pode ser aferida por meio de métricas de transferabilidade e de distância entre distribuições de dados extraídas nos rounds iniciais do treinamento, e se essa compatibilidade é capaz de predizer previamente o desempenho final da federação.

1.2.2 Objetivos específicos

- Avaliar a correlação entre a compatibilidade média e pessimitas obtidas nos rounds iniciais do treinamento federado e o desempenho final do modelo global.
- Avaliar a sensibilidade da compatibilidade em relação a : datasets, modelos de rede neurais, quantidade de clientes da federação e quantidade de épocas locais de treinamento.

2 Revisão Bibliográfica

2.1 Redes neurais e Aprendizado Federado

2.1.1 Redes Neurais

As redes neurais artificiais são técnicas de aprendizado de máquina inspiradas na inteligência humana, em especial no funcionamento dos neurônios biológicos e em suas sinapses (MONTESINOS-LóPEZ; MONTESINOS-LóPEZ; CROSSA, 2022). Uma rede neural artificial estabelece conexões entre suas unidades de processamento (neurônios artificiais) e gera sua saída a partir da organização e dos pesos dessas conexões(ISLAM; CHEN; JIN, 2019). As *ANNs* (*Artificial Neural Networks*), sobretudo as redes neurais profundas (*Deep Neural Networks*), ao contrário dos algoritmos tradicionais de aprendizado de máquina, têm obtido resultados cada vez mais expressivos em aplicações como reconhecimento de fala, processamento de linguagem natural (NLP), visão computacional e análise de imagens (LIU et al., 2017).

(NODA et al., 2015) propôs um sistema que utiliza Deep Learning para reconhecimento de fala combinando informações acústicas e visuais. Neste sistema, o autor reuniu um dataset audiovisual japonês com 400 palavras de seis falantes gravadas em áudio e vídeo, gerou uma versão com ruído dos áudios e treinou um modelo para restaurar o som original, treinou uma CNN para predizer a probabilidade de cada fonema a partir de imagens da boca e, por fim, integrou os dois resultados num modelo capaz de ajustar automaticamente a importância das pistas de áudio e de vídeo, garantindo reconhecimento de palavras mais preciso mesmo em ambientes barulhentos.

(SUMMERS, 2017) apresenta, em seu trabalho, como o deep learning promoveu mudanças radicais nos diagnósticos assistidos por computador (CAD) em imagens médicas, reduzindo erros e tornando a interpretação de imagens mais eficiente. As principais áreas de aplicação do CAD em análise de imagens são radiologia, cardiologia e patologia. Uma dificuldade relatada pelo autor em CAD era a escassez de dados massivos para treinamento, devido à natureza do problema, além do grande esforço e tempo necessários para desenvolver algoritmos manuais voltados a tarefas específicas. Com o deep learning, porém, houve sucesso na resolução de problemas complexos, graças à capacidade dos modelos de aprender características relevantes a partir dos dados de treinamento.

2.1.1.1 Arquitetura geral e funcionamento

O funcionamento de uma rede neural parte de um neurônio artificial, que é uma unidade de processamento conforme representada pela figura 2.1 e composto pelos seguintes elementos:

- Entradas
- Pesos
- viés(bias)
- Função de ativação

O vetor de entradas $x1...xd$ é combinado com os pesos $w1...wd$ e com o viés b em uma soma ponderada. O resultado dessa soma torna-se o argumento de uma função de ativação, que produz a saída do neurônio.

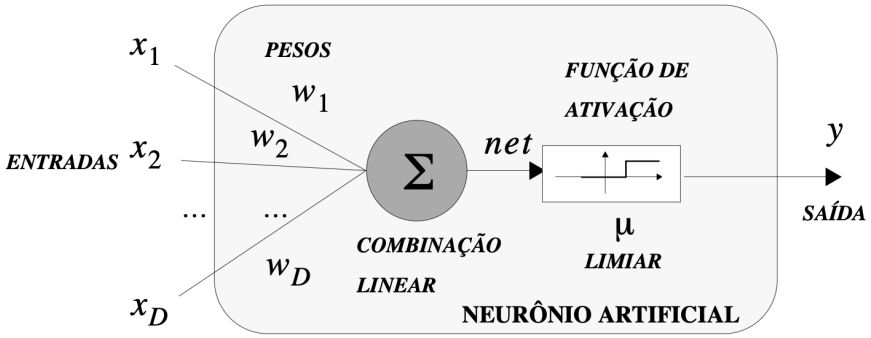

Figura 2.1 – Exemplo de neurônio artificial (adaptado de Costa (COSTA, 2009)).

A figura 2.2 ilustra uma rede neural com multiplas camadas. A camada mais à esquerda(input layer) é a camada de entrada dos dados, a camada mais a direita é a camada de saída(neste exemplo é um neurônio singular), e as camadas do intermediárias são as camadas ocultas(hidden layers).

Em conjunto, esses neurônios formam uma rede neural artificial que mapeia características de entrada em uma saída. Esse aprendizado ocorre camada a camada: cada neurônio ajusta seus pesos para capturar relações entre as características de entrada e a saída desejada (GEORGEVICI; TERBLANCHE, 2019).

2.1.1.2 Função de ativação

As funções de ativação desempenham um papel fundamental em redes neurais, pois permitem ao modelo aprender características abstratas por meio de transformações não lineares. Uma função de ativação linear é aquela que retorna cx como saída, para entradas x e constante c, funcionando como uma identidade. No entanto, a não linearidade é essencial em redes neurais, uma vez que os dados reais normalmente apresentam comportamentos não lineares. Essa propriedade torna o modelo mais flexível e capaz de se ajustar melhor aos dados. Além disso, as funções de ativação devem satisfazer três requisitos principais:

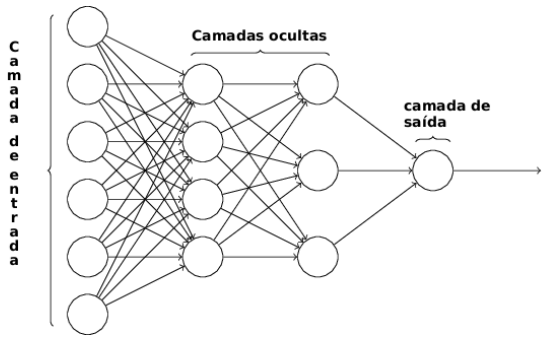

Figura 2.2 – Representação de uma rede neural multi camadas (adaptado de (NIELSEN, 2015, Cap. 1)).

Fonte**:** Nielsen, M. A. *Neural Networks and Deep Learning*. Determination Press, 2015. Licença**:** CC BY-NC 3.0 Unported (<https://creativecommons.org/licenses/by-nc/3.0/>)

- Baixo custo computacional: de forma que não aumentem significativamente o tempo de treinamento ou inferência.
- Fluxo eficiente de gradiente: para evitar problemas de desaparecimento ou explosão de gradientes;
- Preservação da distribuição dos dados:contribuindo para um treinamento estável e rápido da rede.

(DUBEY; SINGH; CHAUDHURI, 2021) . Uma função de ativação muito importante é a função sigmoid :

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{2.1}$$

$$\frac{1}{1+\exp\left(-\sum_j w_j x_j - b\right)} \tag{2.2}$$

desta forma quando o somatório $(-z)$ é um número positivo grande e^{-z} produz um número próximo a zero eo resultado da função fica próximo de um, por outro lado quando $(-z)$ é um número muito negativo, e^{-z} produz um número que tende ao infinito e desta forma a divisão tende a zero(NIELSEN, 2015).

Outra função de ativação notável é a ReLU (Rectified Linear Unit), que foi proposta para contornar a saturação observada na função sigmoid quando as entradas assumem valores muito baixos ou muito altos . Essa saturação provoca o vanishing gradient, situação em que o gradiente da função de custo em relação a um parâmetro torna-se tão reduzido que, ao empregar o gradiente descendente estocástico, quase não ocorrem atualizações desses parâmetros. O vanishing gradient compromete seriamente o treinamento da rede: as camadas mais profundas recebem gradientes ínfimos, o que prejudica a qualidade do aprendizado e impede a rede de convergir adequadamente.

Além disso o calculo de gradiente tem a complexidade computacional alta para essa função. A ReLu é uma função identidade para entradas positivas e 0 para para negativas, como a função varia de $[0,)$, o gradiente para entradas positivas é 1 e 0 para negativas. Desta forma ReLu resolve o problema da complexidade computacional pois o calculo do gradiente é simples, porém ela ainda apresenta o problema do vanishing gradient para inputs negativos. Mesmo com este problema a ReLu e suas variantes apresentam bons resultados (DUBEY; SINGH; CHAUDHURI, 2021).

$$\mathrm{ReLU}(x) \;=\; \max(0,x) \;=\; \begin{cases} x, & x \geq 0, \\ 0, & \text{caso contrário.} \end{cases} \tag{2.3}$$

2.1.1.3 Treinamento

O processo de aprendizado de uma rede neural é sobre encontrar os pesos que fazem a rede exibir a saída desejada(SCHMIDHUBER, 2015) . Para este processo é fornecida uma quantidade significativa de dados chamados dados de treinamento que possuem a entrada e a saída correspondente já pré-definidas, a partir destes dados e das previsões feitas pelo modelo os pesos são ajustados. A figura (2.3) mostra um conjunto de dados com imagens de entrada e o rótulo de saída correspondente.

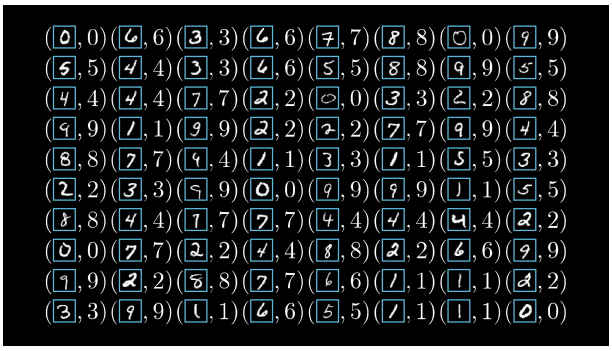

Figura 2.3 – Exemplos de um conjunto de dados de treino (adaptado de (3Blue1Brown,)).

Para dar início ao processo de treinamento os pesos e bias são inicializados de forma aleatória, a rede provavelmente irá prever resultados que divergem muito dos reais. Uma função de perda, pode ser descrita como uma medida de qualidade para a predição da rede neural, de forma mais simples os erros entre a predição e os resultados verdadeiros são avaliados pelo conjunto de treinamento(LI; DOROSLOVACKI; LOEW, 2020). Existem diversas funções de perda descritas na literatura, e para escolhermos a adequada ao nosso modelo é necessário analisar que tipo de tarefa se deseja resolver, como classificação ou regressão (TERVEN et al., 2025) . Neste trabalho, por se tratar de uma tarefa de classificação, é utilizada a função de perda de entropia cruzada.

Entropia cruzada(*Cross entropy*) : é uma função de perda para classificação, também conhecida como perda logarítimica multiclasse, $y_i = [y_{i,1}, \ldots, y_{i,C}] \in \{0,1\}^C$ é o vetor one-hot que indica a classe verdadeira do i-ésimo exemplo (exatamente um $y_{i,j} = 1$ e os demais são zero). e $\hat{p}i$ representa a probabilidade predita pelo modelo do item pertencer a classe. Esse somátorio representa a média da função de perda em todas as amostras do conjunto de dados.

$$L_{\text{CE}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} y_{i,j}\,\log\left(\hat{p}_{i,j}\right) \tag{2.4}$$

2.1.1.4 Gradiente descendente

É o algoritmo mais popular para otimizar redes neurais artificiais, o gradiente descendente visa minimizar uma função de perda $J()$, o gradiente $\nabla_\theta J(\theta)$ é um vetor que aponta para onde a função aumenta. Então, em cada etapa, ajustamos nossos parâmetros θ no sentido a diminuir a inclinação desse vetor. A variável η representa a taxa de aprendizado (learning rate) que é o tamanho dos passos que tomamos para alcançar o mínimo local, este processo é repetido várias vezes até encontrarmos o mínimo local(RUDER, 2016).

$$\theta \leftarrow \theta - \eta\nabla_\theta J(\theta) \tag{2.5}$$

O ADAM (*Adaptive Moment Estimation*) é um algoritmo de otimização amplamente utilizado no treinamento de redes neurais artificiais. Assim como o gradiente descendente, o ADAM tem como objetivo minimizar uma função de perda $J(\theta)$, porém ele o faz utilizando estimativas adaptativas dos momentos de primeira e segunda ordem do gradiente, o que torna o processo de otimização mais estável e eficiente (KINGMA; BA, 2015).

Backpropagation:

O *backpropagation* é um algoritmo para treinar redes neurais, no qual se aplica o gradiente descendente em conjunto com a regra da cadeia para computar os gradientes em todas as camadas da rede (NIELSEN, 2015). Como a rede ainda não passou pelo processo de treinamento, os pesos e os vieses encontram-se inicialmente aleatórios, sendo então definida uma função de custo, responsável por quantificar o erro na camada de saída.

O *forward* é a etapa em que a rede neural, dada uma amostra de entrada, calcula suas ativações até produzir uma saída. Esse processo pode ser descrito pelas seguintes equações:

$$z^l = w^l a^{l-1} + b^l, \qquad a^l = \sigma(z^l), \tag{2.6}$$

onde a^l representa o vetor de ativações da camada l, z^l corresponde à soma ponderada dos neurônios dessa camada, e $\sigma(\cdot)$ é a função de ativação. Observa-se que a ativação de uma camada depende diretamente das ativações da camada anterior, permitindo a propagação das informações pela rede.

A partir da saída da rede, inicia-se o processo de retropropagação do erro. Define-se o termo δ_j^l como o erro do j^{th} neurônio da l^{th} camada. O objetivo do backpropagation é calcular as derivadas parciais da função de custo C em relação aos pesos e vieses. O erro na camada de saída é dado por:

$$\delta^L = \nabla_a C \odot \sigma'(z^L), \tag{2.7}$$

em que $\nabla_a C$ representa o vetor das derivadas parciais do custo em relação às ativações da camada de saída e $\sigma'(z^L)$ corresponde às derivadas da função de ativação.

Para as camadas internas, o erro é propagado recursivamente segundo:

$$\delta^l = \left((w^{l+1})^T \delta^{l+1}\right) \odot \sigma'(z^l), \tag{2.8}$$

permitindo trazer o erro da camada seguinte para a camada atual. Essa equação aplica a regra da cadeia para distribuir corretamente o gradiente pelas camadas ocultas da rede.

A partir do termo de erro δ^l, obtêm-se as taxas de variação da função de custo em relação aos parâmetros:

$$\frac{\partial C}{\partial b^l} = \delta^l, \qquad \frac{\partial C}{\partial w^l} = a^{l-1}\delta^l. \tag{2.9}$$

Esses gradientes são então utilizados na atualização dos pesos e vieses por meio do gradiente descendente,

$$\theta \leftarrow \theta - \eta \nabla_\theta C, \tag{2.10}$$

repetindo-se esse ciclo de *forward*, retropropagação do erro e atualização dos parâmetros até que um critério de parada seja satisfeito.

2.1.2 Redes Neurais Convolucionais

Redes neurais convolucionais (do inglês *convolutional neural networks*, CNN) são uma das arquiteturas mais utilizadas em deep learning, especialmente em visão computacional, mas não se limitam a essa área.

As CNNs são redes de propagação direta capazes de capturar características dos dados por meio de operações convolucionais. Ao projetar uma CNN, deve-se atentar aos seguintes componentes:

- Camada de convolução: etapa de extração de características. A saída dessa operação é denominada *mapa de características*.
- Filtro (*kernel*): pequena matriz cujos coeficientes operam sobre as vizinhanças da matriz de entrada.
- Preenchimento (*padding*): técnica que evita a perda de informações nas bordas, adicionando valores (por exemplo, zeros) ao redor da entrada.

- Passo (*stride*): controla o deslocamento do kernel. Stride=1 aplica a convolução em todas as posições válidas; strides maiores reduzem a densidade de aplicação, diminuindo o custo computacional e o tamanho do mapa de características.
- Amostragem espacial (*pooling*): reduz redundâncias após a convolução. As técnicas mais comuns são *max pooling* (amostragem máxima) e *average pooling* (amostragem média).

Os pesos dos kernels são aprendidos durante o treinamento, permitindo que a rede identifique as características mais relevantes para a tarefa. (LI et al., 2021)

A figura 2.4 ilustra o processo de convolução: dado uma matriz de entrada ela é preenchida para se adequar ao tamanho do filtro, ocorre o processo de convolução e depois o de amostragem.

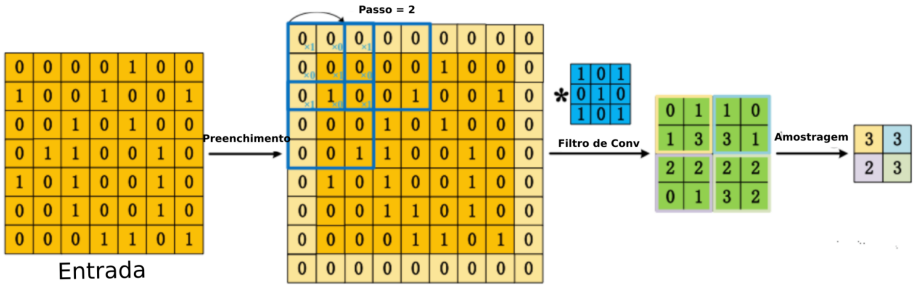

Figura 2.4 – Processo de convolução (adaptado de (LI et al., 2021)).

A figura 2.5 ilustra a arquitetura completa de uma pequena cnn, desde a matriz de entrada, passando pelas diversas camadas de convolução e amostragem, até a camada de saída com o resultado inferido.

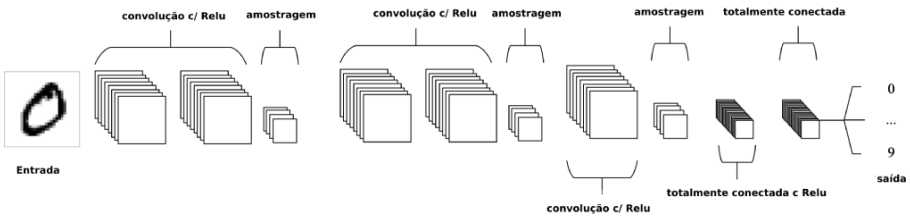

Figura 2.5 – Arquitetura de uma pequena CNN (adaptado de (O'SHEA; NASH, 2015)).

2.1.3 Redes Neurais em Aprendizado Federado

O conceito de aprendizado federado foi introduzido por (MCMAHAN et al., 2017) como uma estratégia para treinar redes neurais a partir de dados descentralizados. Essa abordagem busca solucionar o desafio de utilizar dados sensíveis no treinamento de modelos de aprendizado

de máquina, ao mesmo tempo em que preserva a privacidade das fontes que os geraram. Propôs-se, então, uma alternativa na qual os dados permanecem nos locais em que foram gerados, e o treinamento é realizado por meio de um modelo compartilhado ao qual se agregam as computações locais. Além de preservar a privacidade, essa abordagem permite aproveitar a capacidade computacional de diversos dispositivos em larga escala, viabilizando a execução de projetos robustos que seriam onerosos ou inviáveis se realizados de forma centralizada.(MCMAHAN et al., 2017) A etapa de treinamento no aprendizado federado opera da seguinte maneira: cada dispositivo participante da rede federada é denominado cliente e mantém localmente seus dados de treinamento, que não são compartilhados. Há também um servidor coordenador, que não tem acesso aos dados dos clientes e detém o modelo global de aprendizado de máquina. Em cada rodada, cada cliente atualiza o modelo global com base em seus dados locais, enviando apenas essas atualizações ao servidor. Em seguida, o servidor agrega as atualizações de todos os clientes (por exemplo, por média) para gerar um novo modelo global. Os clientes comunicam-se periodicamente com o servidor para repetir esse processo, completando-se um round cada vez que um novo modelo global é gerado. As rodadas se repetem até que seja atendido um critério de parada estabelecido (YANG et al., 2019).

Figura 2.6 – Esquema de uma rede federada (adaptado de (YANG et al., 2019)

)

A figura 2.6 descreve o comportamento de uma rede federada, no exemplo acima vários clientes(celulares) , executam uma atualização local e retornam essa atualização ao servidor coordenador que gera um novo modelo global e o redistribui.

2.1.3.1 Taxionomia no FL

Segundo (BAO; GUO, 2021) as principais configurações de aprendizado federado podem ser divididas como as seguintes :

- Aprendizado federado vertical: É mais conveniente quando muitos dados são iguais porém as caractéristicas deles são diferentes. Por exemplo , uma universidade e o hospital dessa cidade fizessem uma colaboração, apesar de muitas pessoas estarem presentes nos dois clientes as características dos dados coletados por cada instituição se difere.

- Aprendizado federado horizontal: É a união dos dados, é aplicável quando a maioria das caractéristicas entre clientes da federação são iguais, os dados de cada cliente se diferem, mas eles coletam o mesmo tipo de informação. Um exemplo seria dados sobre alunos de diferentes universidades, cada universidade(cliente) teria diferentes alunos, mas as caractéristicas coletadas seriam parecidas(nome,cpf,idade,matrícula). Desta forma é possivel compartilhar os dados para criar um modelo mais robusto.

- Federated transfer learning: É utilizado quando as amostra e as características das amostras são iguais, desta forma é possível que um cliente transfira conhecimento para outro. Por exemplo uma empresa nova de finanças pode utilizar base de dados públicas para aprender e melhorar seu produto.

2.1.3.2 Formulação

O aprendizado federado pode ser descrita pela seguinte função objetivo:(YANG et al., 2019)

$$\min F(w) := \sum_{k=1}^{m} p_k F_k(w) \tag{2.11}$$

Desta forma : m representa o número total de clientes, k representa um cliente da federação, pk representa o impacto Impacto do cliente k, este impacto é medido pela quantidade de amostrar totais no problema de forma que $pk = 1/n$ ou $pk = nk/n$, em que nk é o numero de amostrar disponiveis para o cliente k e n é o número total de amostras. Fk: função objetivo local para o k-ésimo cliente e é descrita como o risco empírico sobre os dados locais, é a função de perda para cada cliente e nk representa a quantidade de amostras no cliente,

$$F_k(w) = \frac{1}{n_k} \sum_{j=1}^{n_k} f_j^{(k)}\left(w;\, x_j^{(k)},\, y_j^{(k)}\right) \tag{2.12}$$

2.1.3.3 FedAvg

O FedAvg é um dos algoritmos pioneiros e mais utilizados em aprendizado federado. Nesse método, o servidor central coordena o treinamento, enquanto as operações de otimização são executadas localmente pelos clientes, por meio de técnicas como o SGD (gradiente descendente estocástico) (MCMAHAN et al., 2017). O servidor central armazena o modelo global wt, onde t indica a rodada de comunicação.

O algoritmo FedAvg tem 5 hiperparametros :

- C : fração de clientes selecionados para o treinamento.

- B : tamanho do mini-batch local.

- E: Número de épocas locais.
- η: Taxa de aprendizado (learning rate).
- λ: taxa de decaimento de aprendizado.

O algoritmo 2.1 inicia com o servidor configurando o modelo global com parametros aleatórios na rodada $\hat{w}^0$. Em cada rodada t, o servidor escolhe um número de clientes(existem diversas maneiras de fazer a seleção de clientes, uma fração aleatória ou todos os clientes disponíveis são exemplos) e distribui o modelo atual para estes clientes, que passam a tomar o modelo w0 como seu modelo local. Cada cliente então treina seus dados utilizando SGD por Eépocas, e atualizam seu modelo local . Por fim, todos os clientes envolvidos enviam suas versões treinadas ao servidor, que realiza uma agregação das versões por meio de uma média ponderada, em que o impacto de cada cliente é relativo a quantidade de dados que ele possui, resultando em um novo modelo global.

Algoritmo 2.1: Federated Averaging (FedAvg)

Input: Número total de clientes K; fração de clientes por rodada C; número de épocas locais E; taxa de aprendizado η; tamanho do lote B

Output: Pesos globais treinados w

Função `AtualizacaoCliente(`*cliente* k*, pesos* w`):`
    `// Executado em cada cliente selecionado`
    $\mathcal{B} \leftarrow$ dividir o conjunto de dados local P_k em lotes de tamanho B;
    for *cada época local* $i = 1$ *até* E **do**
        for *cada lote* $b \in \mathcal{B}$ **do**
            $w \leftarrow w - \eta \nabla \ell(w; b)$; `// Cálculo do gradiente e atualização dos pesos`
    return w;

`/* Execução no Servidor Central */`
Inicializar os pesos do modelo global w^0;
for *cada rodada de comunicação* $t = 0, 1, 2, \ldots$ **do**
    $m \leftarrow \max(\lfloor C \cdot K \rfloor, 1)$;
    $S_t \leftarrow$ (Selecionar um conjunto aleatório de m clientes);
    for *cada cliente* $k \in S_t$ ***em paralelo*** **do**
        $w_k^{t+1} \leftarrow \text{AtualizacaoCliente}(k, w^t)$;
    `// Agregação dos pesos no servidor`
    $w^{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{n} w_k^{t+1}$; `// onde` n_k `é o tamanho do dataset do cliente` k`, e` $n = \sum_k n_k$

2.1.3.4 Limitações para dados heterogêneos

A distribuição dos dados numa federação pode ser definida como IID se a partir de um dataset qualquer as amostras são independentes e distribuidas de forma igual entre os clientes dessa federação. Non-IID é uma distribuição que por outrora apresenta dependencia entre as amostras e sua distribuição não apresenta caráter uniforme. Algumas das limitações produzidas por dados com caráter Non-IID :

- Performance pior do modelo.
- O modelo pode não convergir com poucos participantes.
- O modelo pode não convergir com um número alto de épocas.

A figura 2.7 exemplifica a diferença entre as duas distribuições de dados.

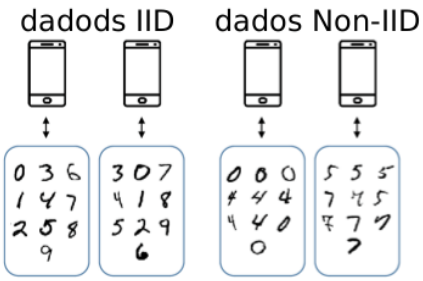

Figura 2.7 – Comparação conjuntos de dados IID e Non-IID (adaptado de (IYER, 2025)).

Pode-se observar três tipos de assimetrias em dados Non-IID:

- assimetria de rótulos: se refere a uma assimetria relativa a quantidade de rótulos ou classe em cada cliente.
- assimetria de características: se refere a uma assimetria na distribuição das características dos dados entre os cliente da federação. Por exemplo alguns clientes podem possuir dados mais ruidosos que outros.
- assimetria de quantidade: se refere a uma assimetria em que cada cliente recebe uma quantidade de dados diferente.

(IYER, 2025).

A 2.8 ilustra a dificuldade que dados Non-IID representam para a convergência de um modelo de aprendizado federado, neste exemplo θ_t representa o modelo global, θ_t^{avg} representa a média dos modelos dos clientes θ_t^1 e θ_t^2.

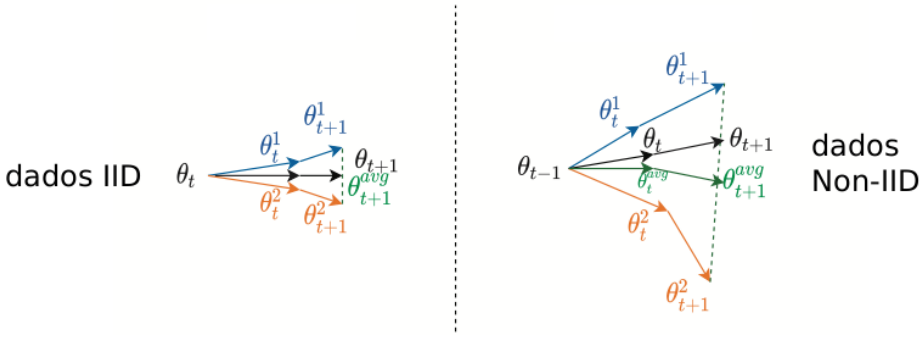

Figura 2.8 – Non-IID x IID (adaptado de (ZHU et al., 2021)).

2.2 Compatibilidade entre clientes e federação com indicadores de transferabilidade

Na seção anterior foram discutidas as limitações impostas por distribuições Non-IID no contexto do aprendizado federado, diante desse cenário é de extrema valia dispor de métodos capazes de estimar o desempenho do modelo durante o treinamento. Este trabalho propoe uma forma de medir a compatibilidade entre modelos treinados em diferentes clientes de uma federação, nesse contexto o conceito de transferabilidade é importante, pois serve como ferramenta para análise destes modelodes.

2.2.1 Métricas de Transferabilidade em FL Horizontal

(BAO et al., 2022) definem que transferir o aprendizado de um modelo de rede neural para outro é um processo comum quando estamos trabalhando com tarefas parecidas. A demanda por supervisão ao se adotar essa prática é menor pois são necessários menos dados rotulados para treinar um novo modelo, se tornando assim um grande facilitador. Um exemplo de transferência de sucesso é o fato de que alguns conjuntos de dados para análises de imagens médicas mesmo possuindo poucos dados rotulados, conseguiram utilizar modelos de redes neurais convolucionais treinadas no dataset *ImageNet* como fonte para transferir conhecimento para o seu modelo e conseguir atingir um bom desempenho na sua tarefa. Uma indagação importante : dado um modelo x , e um modelo y que resolveu uma tarefa similar a que x irá resolver, como estimar se y conseguirá transferir seu conhecimento para x afim de que o mesmo resolva sua tarefa com sucesso. Esta indagação gerou um novo conceito chamado de transferabilidade.

No contexto do FL horizontal, o conceito de transferabilidade pode ser reinterpretado como uma medida de compatibilidade entre modelos locais treinados em distribuições distintas de dados. Diferente do modelo de aprendizado de máquina tradicional, no FL cada cliente contribui indiretamente para o modelo global, assim estimar a transferabilidade entre modelos locais e o modelo global torna-se equivalente a estimar o impacto da contribuição de cada cliente na

federação.

2.2.2 Similaridade dos dados

Métricas como similaridade de cosseno, distância euclidiana e distância de Manhattan são amplamente utilizadas para comparar instâncias representadas como vetores em espaços métricos. Conforme a classificação apresentada por (LEVY; SHALOM; CHALAMISH, 2024), essas medidas integram famílias fundamentadas em operações geométricas no espaço vetorial, ao passo que divergências como Kullback–Leibler e Jensen–Shannon pertencem a famílias especificamente formuladas para comparar distribuições de probabilidade.

No contexto do aprendizado federado, em que os dados frequentemente seguem distribuições não-IID, torna-se pertinente empregar métricas capazes de quantificar discrepâncias estatísticas entre distribuições. Dessa forma, optou-se pela utilização de métricas como KL, Jensen–Shannon, FID e MMD, por sua adequação à análise de divergências entre distribuições

2.2.2.1 Divergência de Kullback-Leibler

A divergência de Kullback-Leibler (KULLBACK; LEIBLER, 1951), permite avaliar o quanto uma distribuição probabilística difere de outra, quando esta é utilizada como uma aproximação. Ao utilizar a divergência é possível quantificar a perda de informação ao substituir uma distribuição probabilística por outra.

$$\mathrm{D}_{\mathrm{KL}}(P \parallel Q) = \sum_{x} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{2.13}$$

A divergência de Kullback-Leibler(KL) de P em relação a Q pode ser definida da seguinte forma:

- $P(x)$: probabilidade do evento x segundo a distribuição P.
- $Q(x)$: probabilidade do evento x segundo a distribuição Q.
- $\log \left(\frac{P(x)}{Q(x)} \right)$ mede o quão diferente é a probabilidade de x em P e Q.
  - É importante ressaltar que se a divisão for 1, o resultado do log é 0.
  - $P(x) > Q(x)$ o log é positivo, Q subestima x.
  - $P(x) < Q(x)$ o log é negativo, Q superestima x.
- $P(x)$ atua como peso ao multiplicar o resultado do $\log$, cada termo é ponderado com sua probabilidade de acontecer em P.

2.2.2.2 Divergência de Jensen-Shannon

JS(LIN, 1991) é uma métrica de distância de distribuições. Segundo(NIELSEN, 2025), a divergência de Jensen-Shannon pode ser interpretada como uma versão simétrica da divergência de Kullback-Leibler, construída a partir da comparação de cada distribuição com uma distribuição intermediária.

$$\mathrm{JS}(P,Q) = \frac{1}{2}\,\mathrm{KL}(P \parallel M) + \frac{1}{2}\,\mathrm{KL}(Q \parallel M)\,, \quad M = \frac{1}{2}(P+Q). \tag{2.14}$$

- P: probabilidade seguindo a distribuição P.
- Q: probabilidade seguindo a distribuição Q.
- M : é a mistura das distribuições.
- KL : É a divergência de Kullback-Leiber

2.2.2.3 Maximum mean discrepancy

A discrepância média máxima é uma métrica estatística que foi proposta por (GRETTON et al., 2012) para avaliar se duas amostras pertecem a mesma distribuição probabilística.

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] = \sup_{f \in \mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(y_i) \right). \tag{2.15}$$

A formulação da *Maximum Mean Discrepancy* (MMD) pode ser definida a partir de uma classe de funções $\mathcal{F}$, em que X e Y representam amostras independentes extraídas de duas distribuições de probabilidade distintas. A MMD corresponde ao supremo, sobre todas as funções pertencentes a $\mathcal{F}$, da diferença entre as médias obtidas pela aplicação de uma função f às amostras de X e Y, medindo, assim, a maior discrepância possível entre as duas distribuições segundo a classe de funções considerada. O operador de supremo garante que a MMD corresponda à maior discrepância possível entre as distribuições, ao selecionar a função em $\mathcal{F}$ que maximiza a diferença entre suas médias empíricas.

Ao considerar essa classe de funções em um espaço de Hilbert, podemos reformular a definição da MMD. Ao assumirmos que $\mathcal{F}$ corresponde ao conjunto de funções pertencentes a um espaço de Hilbert com núcleo reprodutor, torna-se possível representar essas funções implicitamente por meio de um kernel. Dessa forma a maximização realizada pelo supremo deixa de ser efetuada diretamente sobre as funções f, passando a ser determinadas pelas propriedades geométricas induzidas pelo kernel.

Sob essa perspectiva geométrica, cada distribuição de probabilidade pode ser representada como um ponto no espaço de Hilbert, obtido a partir da média das representações de suas amostras. A MMD passa, então, a corresponder à distância entre esses pontos, sendo a função que atinge o supremo aquela alinhada à direção que conecta as representações das duas distribuições

nesse espaço. Dessa forma, o kernel define implicitamente um espaço no qual a maximização envolvida na definição da MMD é garantida teoricamente, permitindo calcular a discrepância entre distribuições sem a necessidade de identificar explicitamente a função que a maximiza. Entre as possíveis escolhas de kernel, opta-se pelo kernel Gaussiano, o qual induz um espaço de Hilbert suficientemente rico, capaz de capturar diferenças sutis entre distribuições distintas.

A definição do kernel Gaussiano

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \tag{2.16}$$

$$\mathrm{MMD}^2(X, Y) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, y_j), \tag{2.17}$$

Diante das definições anteriores, a MMD com kernel pode ser interpretada como a soma das médias do kernel aplicado às amostras pertencentes ao mesmo conjunto, tanto em X quanto em Y, subtraída da média do kernel aplicado entre amostras de conjuntos distintos. Desta forma, o primeiro termo quantifica a similaridade média entre as amostras do conjunto X, o segundo termo mede a similaridade média entre as amostras do conjunto Y, enquanto o terceiro termo captura a similaridade entre amostras de X e Y. Todos esses termos são calculados a partir de um mesmo kernel, o qual define a noção de similaridade no espaço de Hilbert induzido e permite quantificar a discrepância entre as distribuições a partir da comparação entre similaridades intra-distribuição e inter-distribuição.

2.2.2.4 Fréchet Inception Distance (FID)

GANs (*generative adversarial networks*) (GOODFELLOW et al., 2014) são modelos capazes de gerar imagens realistas e produzir texto com grande sucesso. Essas arquiteturas são compostas por duas redes neurais: uma rede geradora, responsável por criar dados sintéticos a partir de variáveis aleatórias, e uma rede discriminadora, responsável por distinguir dados reais de dados sintéticos. Desta forma, existe uma espécie de jogo minimax entre essas duas redes onde a geradora tenta gerar amostras que a discriminadora não consegue distinguir se são sintéticas ou não.

A Fréchet Inception Distance (FID)(HEUSEL et al., 2017) entra como métrica de avaliação para determinar se o conjunto de imagens geradas se assemelha, em nível de distribuição estatística, ao conjunto de imagens reais. Formalmente, a FID é definida da seguinte forma:

$$\mathrm{FID} = d^2\big((m, C), (m_w, C_w)\big) = \|m - m_w\|_2^2 + \mathrm{Tr}\Big(C + C_w - 2(CC_w)^{\frac{1}{2}}\Big). \tag{2.18}$$

Nessa definição, (m, C) representam, respectivamente, a média e a matriz de covariância das features extraídas das imagens sintéticas, enquanto (m_w, C_w) representam a média e a matriz

de covariância das features extraídas das imagens reais. Essas features são obtidas a partir da camada de codificação de um modelo Inception pré-treinado. O termo $d^2(\cdot,\cdot)$ denota a distância de Fréchet ao quadrado entre duas distribuições Gaussianas multivariadas.

O termo $\|m - m_w\|_2^2$ mede a diferença entre as médias das distribuições. Já o termo $\mathrm{Tr}\Big(C + C_w - 2(CC_w)^{\frac{1}{2}}\Big)$ avalia a discrepância entre as matrizes de covariância das distribuições real e sintética.

2.2.3 Compatibilidade entre clientes

No contexto do aprendizado de máquina, a transferabilidade entre conjuntos de dados pode ser avaliada por meio de medidas de similaridade ou distância, as quais buscam quantificar o quão bem o conhecimento aprendido em um domínio pode ser reutilizado em outro (WANG et al., 2025). Em cenários de aprendizado federado, essa noção torna-se particularmente relevante devido à heterogeneidade dos dados distribuídos entre os clientes.

Dessa forma, introduzimos o conceito de *c*ompatibilidade entre clientes, cujo objetivo é mensurar, por meio de indicadores de transferabilidade, tais como a Fréchet Inception Distance (FID), a Maximum Mean Discrepancy (MMD) e divergências estatísticas clássicas, incluindo Kullback-Leibler (KL) e Jensen-Shannon (JS), o quão adequados diferentes conjuntos de clientes são para colaborar em uma mesma federação. Esses indicadores são calculados a partir da comparação entre modelos locais e o modelo global nos rounds iniciais do treinamento federado.

Intuitivamente, a compatibilidade entre clientes busca quantificar o grau em que os conhecimentos aprendidos localmente podem ser integrados de maneira coerente no processo federado. Uma federação com alta compatibilidade tende a apresentar atualizações locais que não introduzem discrepâncias estatísticas excessivas em relação ao modelo global, favorecendo uma agregação estável e uma convergência adequada do modelo ao longo das rodadas. Por outro lado, baixos níveis de compatibilidade podem indicar a presença de distribuições significativamente divergentes, potencialmente prejudicando a estabilidade e o desempenho final do modelo global.

2.2.4 Formulação matemática da compatibilidade entre clientes

Considere uma federação composta por K clientes, treinada ao longo de rodadas $t = 0, 1, 2, \ldots, T$. Seja $\mathbf{w}^{(t)}$ o modelo global na rodada t e $\mathbf{w}_k^{(t)}$ o modelo local do cliente k ao final da rodada t. Denote por $\mathcal{D}_k = \{x_i^{(k)}\}_{i=1}^{n_k}$ o conjunto de amostras disponíveis no cliente k, contendo n_k instâncias.

Extração de representações

Seja $\phi(\mathbf{w}, x)$ uma função que representa a saída intermediária da rede neural (parametrizada por $\mathbf{w}$) para uma amostra x. Essa função pode corresponder aos *logits* ou às *features* de uma camada interna da rede, sem necessariamente alcançar a última camada de classificação.

Para cada cliente k, definem-se duas coleções de representações empíricas:

- Representações induzidas pelo modelo local no round 2:

$$\mathcal{Z}_k^{\text{local}} = \left\{ \phi\left(\mathbf{w}_k^{(2)}, x\right) \;\middle|\; x \in \mathcal{D}_k \right\}$$

- Representações induzidas pelo modelo global do round 1:

$$\mathcal{Z}_k^{\text{global}} = \left\{ \phi\left(\mathbf{w}^{(1)}, x\right) \;\middle|\; x \in \mathcal{D}_k \right\}$$

Observa-se que ambas as distribuições são construídas sobre o mesmo conjunto de dados $\mathcal{D}_k$, garantindo comparabilidade direta entre os modelos.

Compatibilidade por cliente

Seja $d(\cdot, \cdot)$ uma métrica de divergência ou discrepância entre distribuições, tal como Fréchet Inception Distance (FID), Maximum Mean Discrepancy (MMD), divergência de Kullback-Leibler (KL) ou divergência de Jensen-Shannon (JS).

Define-se a compatibilidade do cliente k na rodada 2 como:

$$C_k^{(2)} = d\Big(\mathcal{Z}_k^{\text{local}}, \mathcal{Z}_k^{\text{global}}\Big)$$

ou, explicitando a dependência nos modelos:

$$C_k^{(2)} = d\Big(\phi(\mathbf{w}_k^{(2)}, \mathcal{D}_k), \phi(\mathbf{w}^{(1)}, \mathcal{D}_k)\Big)$$

Essa quantidade mede o grau de discrepância estatística entre as representações produzidas pelo modelo local atualizado e aquelas produzidas pelo modelo global da rodada anterior, avaliadas sobre os dados do próprio cliente.

Compatibilidade média da federação

A compatibilidade média da federação na rodada 2 é definida como:

$$\bar{C}^{(2)} = \frac{1}{K} \sum_{k=1}^{K} C_k^{(2)}$$

Essa medida sintetiza o nível médio de discrepância entre os modelos locais e o modelo global da rodada anterior, considerando todos os clientes da federação.

Compatibilidade pessimista

Além da medida média, pode-se definir uma versão baseada no pior caso, capturando o nível máximo de discrepância observado entre os clientes:

$$C_{\max}^{(2)} = \max_{k \in \{1,\ldots,K\}} C_k^{(2)}$$

Essa definição enfatiza o impacto potencial de clientes altamente heterogêneos, os quais podem introduzir instabilidades no processo de agregação e comprometer a convergência do modelo global.

Ganho global do modelo

Seja $A^{(t)}$ a acurácia global do modelo na rodada t, avaliada sobre o conjunto de teste global.

Define-se o ganho global total do experimento como:

$$\Delta A = A^{(T)} - A^{(0)}$$

onde T representa a rodada final do treinamento.

O valor ΔA quantifica a melhoria total do modelo global ao longo de todo o processo federado.

3 Desenvolvimento

Neste capítulo são apresentados os métodos utilizados para a realização do trabalho, e este se organiza da seguinte forma : inicialmente é apresentado como medir a compatibilidade entre clientes 3.1, em seguida é apresentado o ambiente federado que irá realizar os experimentos 3.2, descreve-se também os conjuntos de dados utilizados 3.3, os modelos de rede neural utilizados e suas configurações 3.4,e por último a configuração dos experimentos 3.5.

3.1 Medição da compatibilidade entre clientes preservando a privacidade

A hipótese central deste trabalho basea-se em medir as diferenças entre os dados dos clientes de uma federação por meio de métricas de transferabilidade. Após a realização dessas medições, calcula-se a correlação entre a média ou o pior caso dos indicadores de transferabilidade nos rounds iniciais da federação e o desempenho final do modelo global. A partir dessa relação, define-se a compatibilidade entre os clientes.

Entretanto, para mensurar diretamente a diferença entre as distribuições de dados de cada cliente, seria necessário ter acesso a essas distribuições, o que viola diretamente o princípio fundamental de privacidade do aprendizado federado. Diante dessa limitação, adotou-se uma estratégia inspirada na proposta de (YASHWANTH et al., 2024), que apresenta um método adaptativo baseado na divergência de Kullback-Leibler para lidar com a heterogeneidade dos dados durante o treinamento federado.

Diferentemente do trabalho citado,o objetivo desta pesquisa não é propor um mecanismo para a melhoria do desempenho do aprendizado federado, mas sim investigar de forma empírica a possibilidade de prever o sucesso de uma federação a partir de indicadores observados nos estágios iniciais do treinamento.

A estratégia adotada consiste em manter, a cada round de treinamento, uma cópia do modelo global da federação. Cada cliente realiza uma rodada de treinamento local a partir deste modelo global,conforme o protocolo padrão do aprendizado federado.

Ao final de cada round, os dados de cada cliente são utilizados como entrada tanto para o modelo global quanto para o modelo local, ambos operando em modo de avaliação. Nessa etapa, são extraidas as representações internas *features* e os *logits* correspondentes a cada amostra do conjunto de dados do cliente.

Parte-se da premissa que o modelo global atua como uma representação agregada do conhecimento proveniente de todos os clientes da federação, uma vez que seus parâmetros são

atualizados a partir das contribuições locais. Em contraste o modelo local, após a realização de uma rodada de treinamento, tende a refletir de forma mais pronunciada as características específicas dos dados do respectivo cliente.

De posse dos *logits* e das *features* extraidos de cada modelo, aplicam-se as métricas de transferabilidade descritas na seção **??**, as quais quantificam a discrepância entre modelo global e o modelo local por meio de um valor escalar.

Em seguida, calcula-se a média ou o pior caso dos valores de cada métrica de transferabilidade no segundo *round* de treinamento, a qual é posteriormente correlacionada com o desempenho final do modelo global da federação.

A escolha do round 2 justifica-se pelo fato de que, após o primeiro ciclo de agregação, o modelo global já incorpora informações provenientes de todos os clientes da federação. No segundo round, por sua vez, o modelo local passa a refletir de maneira mais acentuada as particularidades dos dados de cada cliente . Dessa forma, a comparação entre o modelo global resultante do primeiro round e o modelo local após o segundo round permite capturar discrepâncias relevantes entre os clientes.

3.2 Configuração do ambiente federado

A linguagem de programação escolhida para realização do trabalho foi *Python 3.0*, pois é uma linguagem que apresenta um rico suporte para diversos algoritmos de *machine learning* , processamento e vizualização de dados. Além de também ser compatível com o *Flower framework*, utilizado para a realização dos experimentos. O ambiente para realização de experimentos foi de computadores com a seguinte configuração: processador intel i9-10900, 128gb de memória ram ddr4, gpu RTX3090 com 24gb de ram.

Posteriormente, para tratamento de dados e análises graficas foi utilizado o *Google Collab*, pois sua configuração e vizualização de resultados é simples e rápida.

A configuração de um sistema de aprendizado federado não é trivial. Comunicação entre cliente e servidor, ambiente compatível de computação entre clientes, distribuição de dados entre os clientes, implementação dos algorítmos federados,etc são alguns dos desafios. (BEUTEL et al., 2020) provê um *framework* que facilita e cordena todo o processo de aprendizado federado, além de possuir bibliotecas com grande variedades de algorítmos e estratégias federadas. Por estes motivos o *Flower* foi escolhido para a realização dos experimentos deste trabalho.

3.3 Conjunto de dados

A escolha dos conjuntos de dados se deu da seguinte maneira: procuramos datasets de imagens que já foram utilizados como *benchmarks* em outras pesquisas de aprendizado de

máquina. Para promover robustês as métricas também utilizamos diferentes conjuntos de dados para analisar a resiliencia da metrica em relação a diferentes conjuntos de dados.

Os conjuntos utilizados foram :

- *CIFAR-10*, proposto por (KRIZHEVSKY, 2009) possui 60.000 imagens coloridas , sendo 50.000 para treinamento e 10.000 para teste. Existem 10 rótulos nesse conjunto sendo eles : avião, automóvel, pássaro, gato, cervo, cachorro, sapo, cavalo, navio, caminhão. A distribuição dos rótulos é simétrica.

- *Fashion-MNIST*, proposto por (XIAO; RASUL; VOLLGRAF, 2017), é um conjunto de dados composto por 70.000 imagens em escala de cinza, com resolução de 28×28 pixels, sendo 60.000 imagens destinadas ao treinamento e 10.000 ao teste. O conjunto possui 10 classes relacionadas a itens de vestuário, incluindo camiseta/top, calça, pulôver, vestido, casaco, sandália, camisa, tênis, bolsa e bota. O *Fashion-MNIST* foi desenvolvido como uma alternativa mais desafiadora ao conjunto MNIST tradicional, mantendo a mesma estrutura e balanceamento entre as classes.

- *Blood-MNIST*, apresentado por (YANG et al., 2021), faz parte da coleção MedMNIST e é composto por imagens microscópicas de células sanguíneas humanas. O conjunto contém aproximadamente 17.092 imagens coloridas com resolução de 28×28 pixels, distribuídas em 8 classes que representam diferentes tipos de células do sangue, como neutrófilos, linfócitos, monócitos, eosinófilos e basófilos. O *Blood-MNIST* apresenta um cenário mais complexo e próximo de aplicações reais, sendo amplamente utilizado como *benchmark* em tarefas de classificação de imagens médicas.

3.4 Arquiteturas de Redes Neurais

A biblioteca *PyTorch* foi utilizada para a implementação dos modelos de redes neurais, bem como para os procedimentos de treinamento, teste e modificações estruturais nas arquiteturas. A escolha dessa biblioteca se deu por sua ampla adoção na comunidade científica, flexibilidade para pesquisa experimental e compatibilidade nativa com o framework de aprendizado federado *Flower* (PASZKE et al., 2019).

Os modelos escolhidos contemplam arquiteturas convolucionais com diferentes níveis de profundidade e complexidade. O primeiro modelo utilizado foi a rede *ResNet-18* (HE et al., 2016), empregada em sua versão pré-treinada no *dataset ImageNet*.

Como segunda arquitetura, foi utilizada uma rede convolucional inspirada na família *VGG* (SIMONYAN; ZISSERMAN, 2014), denominada neste trabalho como *LightCNN*. Diferentemente das versões clássicas da VGG, a arquitetura proposta é substancialmente mais leve, composta por blocos convolucionais sequenciais com filtros de pequena dimensão (3×3),

seguidos por funções de ativação ReLU e operações de *max pooling*. Ao final do extrator de características, emprega-se uma camada de *Global Average Pooling*, reduzindo a dimensionalidade espacial das ativações antes da camada totalmente conectada de classificação.

Essa escolha arquitetural permite reduzir significativamente o número de parâmetros treináveis, tornando o modelo mais adequado a cenários de aprendizado federado, nos quais restrições de comunicação e capacidade computacional dos clientes são fatores relevantes.

O propósito da utilização de múltiplas arquiteturas foi análogo ao adotado na escolha dos conjuntos de dados: investigar se as métricas propostas são robustas e consistentes quando aplicadas a diferentes modelos de redes neurais, variando em profundidade, capacidade de representação e número de parâmetros.

3.5 Configuração dos experimentos

O *Flower Framework* disponibiliza um modo de execução denominado *simulation*, o qual permite a criação de clientes simulados para a realização de experimentos de aprendizado federado sem a necessidade de dispositivos remotos reais. Nesse modo, tanto o servidor coordenador quanto os clientes são executados na mesma máquina, cabendo ao *Flower* gerenciar internamente a troca de mensagens e a orquestração do processo de treinamento.

3.5.1 Particionamento dos dados

Na etapa de particionamento dos dados, buscou-se simular diferentes cenários de federações heterogêneas, de forma a analisar o impacto da não homogeneidade estatística entre os clientes. Para esse fim, foi utilizado o particionador baseado na distribuição Dirichlet, disponibilizado nativamente pelas bibliotecas do *Flower*. Esse particionador permite controlar o grau de heterogeneidade dos dados por meio do parâmetro α, sendo que valores mais elevados de α conduzem a distribuições progressivamente mais próximas do cenário IID.

Em todos os experimentos, mantiveram-se fixos o número de clientes, o número de épocas locais, o número de rodadas federadas, o conjunto de dados e a arquitetura do modelo de rede neural. A variação experimental concentrou-se exclusivamente no parâmetro α, o qual assumiu os valores $\{0.05, 0.2, 0.5, 1.0, 3.0\}$. Esses valores representam, respectivamente, cenários de altíssima heterogeneidade, alta heterogeneidade, heterogeneidade moderada, cenário quase IID e cenário praticamente IID.

3.6 Protocolo de treinamento e cenários experimentais

Nesta seção são descritos os protocolos de treinamento utilizados nos experimentos, bem como os diferentes cenários experimentais considerados neste trabalho. Cada cenário é definido

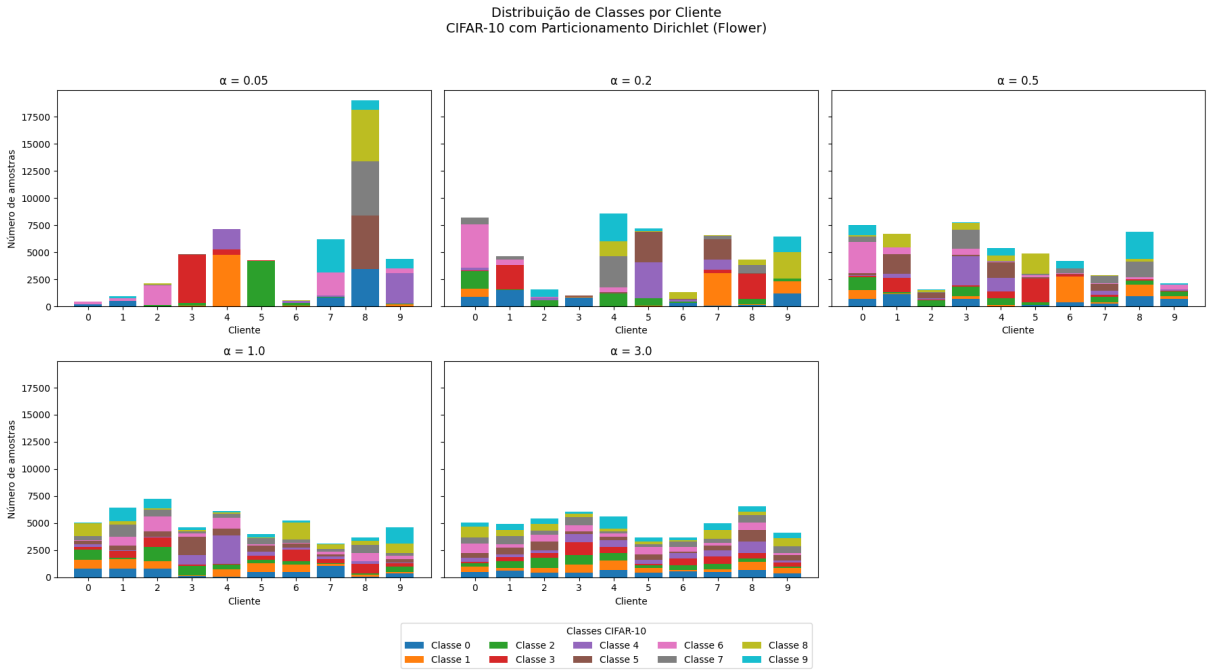

Figura 3.1 – Distribuições de classes e amostra entre clientes para diferentes valores de α .

a partir da combinação entre conjunto de dados, arquitetura de rede neural e número de clientes. Como definido na seção 3.5.1, todos os experimentos compartilham o mesmo conjunto de valores do parâmetro α, responsável por controlar o grau de heterogeneidade dos dados.

Embora os cenários experimentais variem em termos de conjunto de dados, arquitetura e número de clientes, alguns aspectos do protocolo de treinamento são comuns a todos os experimentos.

Em particular, para os conjuntos de dados CIFAR-10 e Fashion-MNIST, que fornecem apenas uma divisão original entre treinamento e teste, o conjunto de treinamento de cada cliente é subdividido localmente em 80% para treinamento e 20% para validação. No caso do conjunto Blood-MNIST, cuja divisão padrão já contempla subconjuntos distintos de treinamento, validação e teste (aproximadamente 70%, 10% e 20%, respectivamente), essas partições originais são respeitadas.

Além disso, as seguintes configurações são mantidas fixas em todos os experimentos realizados:

- A acurácia global de cada *round* é avaliada no conjunto global de teste.
- *Batch size*: 64
- Função de perda: *CrossEntropyLoss*
- Otimizador: *Adam*

- Algoritmo de aprendizado federado: *FedAvg*
- learning rate: 0.001

3.6.1 Experimentos com o conjunto CIFAR-10

Nesta seção apresentamos a configuração dos experimentos que utilizam a base de dados CIFAR-10. A tabela 3.1 detalha os 5 experimentos realizados e suas variações de configurações.

3.6.2 Experimentos com o conjunto Fashion-MNIST

Nesta seção apresentamos a configuração dos experimentos que utilizam a base de dados Fashion-MNIST. A tabela 3.2 mostra os 3 experimentos realizados e suas variações de configurações.

3.6.3 Experimentos com o conjunto Blood-MNIST

Nesta seção apresentamos a configuração dos experimentos que utilizam a base de dados Blood-MNIST. Diferentemente dos demais conjuntos de dados, os experimentos realizados com o Blood-MNIST foram repetidos 30 vezes, com o objetivo de garantir maior robustez e relevância estatística dos resultados obtidos. A tabela 3.3 detalha os 2 experimentos realizados e suas configurações.

Tabela 3.1 – Configurações experimentais utilizando o conjunto CIFAR-10

Exp.	**Clientes**	**Rounds**	**Épocas locais**	**Arquitetura**
1	10	10	10	ResNet-18
2	25	10	10	ResNet-18
3	10	10	1	ResNet-18
4	10	10	10	LightCNN
5	25	10	10	LightCNN

Tabela 3.2 – Configurações experimentais utilizando o conjunto Fashion-MNIST

Exp.	**Clientes**	**Rounds**	**Épocas locais**	**Arquitetura**
6	10	10	10	ResNet-18
7	25	10	10	ResNet-18
8	100	20	10	ResNet-18

Tabela 3.3 – Configurações experimentais utilizando o conjunto Blood-MNIST

Exp.	**Clientes**	**Rounds**	**Épocas locais**	**Arquitetura**
9	10	10	2	ResNet-18
10	25	10	2	ResNet-18

4 Resultados

Nesta seção são apresentadas as correlações de Pearson e Spearman entre as medidas de compatibilidade entre clientes e o ganho global do modelo, formalmente definido como $\Delta A = A^{(T)} - A^{(0)}$.

A compatibilidade é estimada a partir de diferentes indicadores de discrepância estatística, agrupados em duas categorias distintas. O primeiro grupo é composto por métricas baseadas em representações intermediárias da rede neural, isto é, calculadas a partir de *features* extraídas dos modelos: Fréchet Inception Distance (FID) e Maximum Mean Discrepancy (MMD). Essas métricas operam diretamente no espaço vetorial das representações induzidas pelos modelos.

O segundo grupo é composto por métricas baseadas em divergências entre distribuições de probabilidade derivadas dos *logits*: divergência de Kullback-Leibler (KL) e divergência de Jensen-Shannon (JS). Diferentemente das anteriores, essas métricas não dependem explicitamente do espaço de *features*, mas da comparação entre distribuições probabilísticas associadas às saídas dos modelos.

Em todos os experimentos, a variável de interesse é o ganho global ΔA, definido como a diferença entre a acurácia do modelo na rodada final e a acurácia na rodada inicial ($t = 0$). Essa escolha permite isolar a contribuição efetiva do treinamento federado, desconsiderando eventuais ganhos oriundos de pré-treinamento ou da inicialização do modelo.

4.1 Experimentos CIFAR-10

4.1.1 Experimento 1

O Experimento 1 foi conduzido com 10 clientes, 10 rodadas globais e 10 épocas locais por cliente, utilizando a arquitetura ResNet-18 no conjunto CIFAR-10. A Tabela 4.1 apresenta as correlações de Pearson e Spearman entre a compatibilidade média e o ganho global ΔA.

Observa-se que todas as métricas analisadas apresentaram correlação fortemente negativa com ΔA. Esse comportamento indica que maiores níveis de incompatibilidade estatística entre os modelos locais e o modelo global na rodada inicial estão associados a menores ganhos finais de desempenho ao longo do treinamento federado.

As métricas baseadas em representações intermediárias (FID e MMD) exibiram correlações praticamente lineares com ΔA, com coeficientes de Pearson próximos de -1 e coeficientes de Spearman iguais a -1. Esse resultado sugere uma relação monotônica quase perfeita entre o deslocamento no espaço de *features* e o desempenho final da federação.

De maneira consistente, as métricas baseadas em divergências probabilísticas (KL e

JS) também apresentaram correlação negativa forte, com coeficientes de Pearson em torno de $-0{,}8$ e Spearman igual a -1. Embora ligeiramente menos intensas em termos lineares quando comparadas às métricas baseadas em *features*, mantêm o mesmo padrão monotônico de associação com o ganho global.

Como todas as métricas utilizadas são medidas de dissimilaridade, valores mais elevados de compatibilidade indicam maior discrepância estatística entre o modelo local e o modelo global. Assim, o padrão observado reforça a hipótese de que maiores níveis de incompatibilidade nas rodadas iniciais tendem a comprometer o ganho global do modelo federado.

experimento 1

	pearson	spearman
fid	-0.961	-1.000
js	-0.840	-1.000
kl_divergence	-0.822	-1.000
mmd	-0.976	-1.000
delta_global_accuracy	1.000	1.000

Tabela 4.1 – Tabela de correlação entre a compatibilidade media e delta global.

A Tabela 4.2 apresenta os resultados considerando a compatibilidade pessimista, definida como o maior valor da métrica de dissimilaridade observado entre os clientes na rodada 2, isto é, o pior caso em termos de incompatibilidade estatística.

Observa-se que a correlação entre a compatibilidade pessimista e o ganho global ΔA permanece fortemente negativa para todas as métricas analisadas. Os coeficientes de Pearson mantêm-se próximos de -1, enquanto os coeficientes de Spearman são iguais a -1, indicando uma relação monotônica praticamente perfeita.

Esse resultado sugere que, mesmo quando a federação é avaliada sob a perspectiva do cliente mais discrepante, a associação entre incompatibilidade estatística inicial e desempenho final do modelo global permanece consistente. Em outras palavras, a presença de clientes altamente heterogêneos parece estar diretamente relacionada à redução do ganho global ao longo do treinamento federado.

A similaridade entre os resultados obtidos com a compatibilidade média e com a compatibilidade pessimista indica que, neste cenário experimental, tanto o comportamento agregado quanto o pior caso refletem de maneira consistente a dinâmica de convergência do modelo global.

experimento 1

	pearson	spearman
fid	-0.962	-1.000
js	-0.911	-1.000
kl_divergence	-0.915	-1.000
mmd	-0.975	-1.000
delta_global_accuracy	1.000	1.000

Tabela 4.2 – Tabela de correlação entre compatibilide pessimista e delta global.

A imagem 4.1 mostra os gráficos da relação entre a compatibilidade de clientes pelas diferentes métrica de transferabilidade e o delta de acurácia global do modelo federado. Os pontos em azul refletem a medição do experimento feita com 0.05 α , em laranja com α de 0.2, em verde com α de 0.5, em vermelho com α de 1.0, e por último em roxo com α de 3.0.

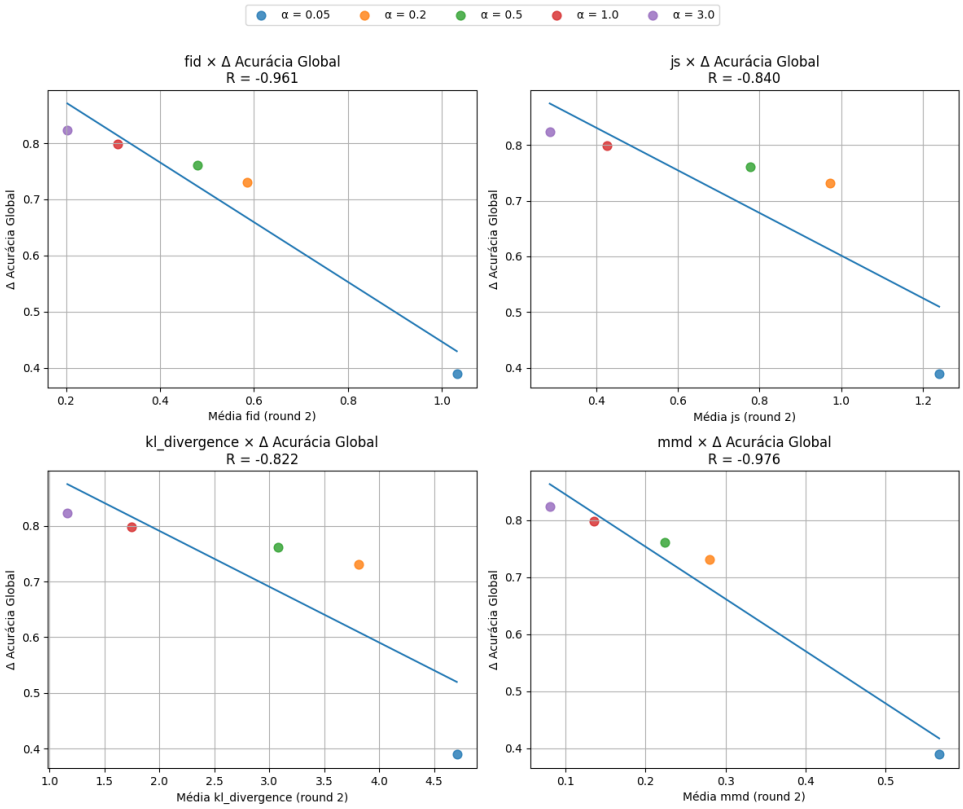

Figura 4.1 – Gráfico que relaciona a compatibilidade media com o delta de acurácia global do modelo, para o experimento 1

4.1.2 Experimento 2

O Experimento 2 mantém a mesma configuração do Experimento 1, conjunto CIFAR-10, arquitetura ResNet-18, 10 rodadas globais e 10 épocas locais ,diferindo apenas no número de clientes, que passa de 10 para 25, conforme apresentado na Tabela 3.1.

experimento 2

	pearson	spearman
fid	-0.909	-1.000
js	-0.730	-0.900
kl_divergence	-0.703	-0.900
mmd	-0.941	-1.000
delta_global_accuracy	1.000	1.000

Tabela 4.3 – Tabela do experimento 2, que correlaciona a compatibilidade média com o delta de acurácia global do modelo

A Tabela 4.3 apresenta as correlações entre a compatibilidade média e o ganho global ΔA. Observa-se que, mesmo com um número de clientes 2,5 vezes maior, todas as métricas mantêm correlação negativa elevada com ΔA, indicando que o aumento da heterogeneidade potencial decorrente do maior número de clientes não altera o padrão geral observado no Experimento 1.

As métricas baseadas em representações intermediárias (FID e MMD) continuam apresentando correlações negativas fortes, com coeficientes de Spearman iguais a -1, indicando relação monotônica perfeita. Embora os coeficientes de Pearson sejam ligeiramente menores em magnitude quando comparados ao experimento anterior, o padrão de associação permanece consistente.

De forma semelhante, as métricas baseadas em divergências probabilísticas (KL e JS) também mantêm correlação negativa significativa com ΔA. Ainda que os coeficientes lineares sejam moderadamente inferiores aos observados para FID, o comportamento monotônico permanece inalterado.

De maneira geral, os resultados indicam que o aumento no número de clientes não altera substancialmente a relação entre compatibilidade inicial e ganho global, sugerindo que o fenômeno observado no Experimento 1 é robusto em relação à variação da cardinalidade da federação.

A Tabela 4.4 apresenta os resultados considerando a compatibilidade pessimista, definida

experimento 2

	pearson	spearman
fid	-0.921	-1.000
js	-0.816	-1.000
kl_divergence	-0.803	-1.000
mmd	-0.844	-1.000
delta_global_accuracy	1.000	1.000

Tabela 4.4 – Tabela do experimento 2,que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo

a partir do maior valor da métrica de dissimilaridade observado entre os clientes na rodada 2.

De maneira consistente com a análise baseada na compatibilidade média, não se observam diferenças substanciais nos coeficientes de correlação. A associação entre compatibilidade e ganho global ΔA permanece fortemente negativa para todas as métricas avaliadas.

As métricas baseadas em representações intermediárias (FID e MMD) continuam apresentando coeficientes de Pearson próximos de -1 e correlação de Spearman igual a -1, indicando relação monotônica praticamente perfeita. As divergências probabilísticas (KL e JS) mantêm o mesmo padrão de comportamento, com correlações negativas elevadas.

A proximidade entre os resultados obtidos pelas abordagens média e pessimista sugere que, neste cenário com 25 clientes, tanto o comportamento agregado quanto o pior caso refletem de maneira equivalente a dinâmica de convergência do modelo global.

No gráfico 4.2 podemos percerber a compatibilidade média no eixo horizontal e o delta de acurácia global no eixo vertical.

4.1.3 Experimento 3

O Experimento 3 mantém a mesma configuração estrutural dos Experimentos 1 e 2 — conjunto CIFAR-10, arquitetura ResNet-18 e 10 rodadas globais , diferindo apenas no número de épocas locais, que neste caso foi reduzido para 1, conforme apresentado na Tabela 3.1.

A Tabela 4.5 apresenta as correlações entre a compatibilidade média e o ganho global ΔA. Observa-se que as métricas mantêm correlação negativa com ΔA, com coeficientes de Pearson variando aproximadamente entre $-0{,}65$ e $-0{,}85$, indicando ainda uma associação linear relevante entre incompatibilidade inicial e desempenho final.

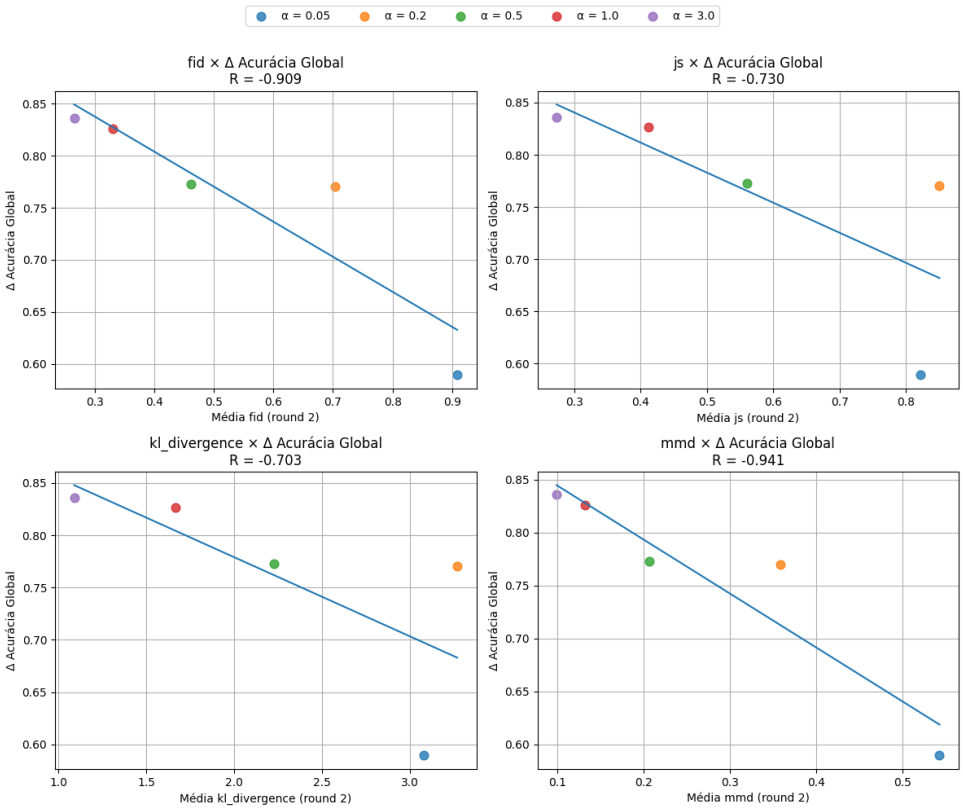

Figura 4.2 – Gráfico que relaciona a compatibilidade média com o delta de acurácia global do modelo, para o experimento 2

Entretanto, diferentemente dos Experimentos 1 e 2, os coeficientes de Spearman apresentam valores próximos de $-0{,}1$, indicando ausência de relação monotônica forte entre as variáveis. Esse resultado sugere que, embora exista tendência linear negativa, a ordenação relativa dos valores de compatibilidade não acompanha de forma consistente a ordenação dos ganhos globais.

A redução no número de épocas locais implica menor afastamento entre os modelos locais e o modelo global a cada rodada, o que pode reduzir a magnitude das discrepâncias estatísticas capturadas pelos indicadores de compatibilidade. Ainda assim, a manutenção de correlação linear negativa indica que a medida proposta preserva capacidade explicativa mesmo em cenários com atualizações locais menos intensas.

Esse experimento é particularmente relevante por aproximar o cenário ao utilizado nos experimentos com o conjunto Blood-MNIST, nos quais também se emprega número reduzido de épocas locais. Dessa forma, os resultados sugerem que o conceito de compatibilidade entre clientes não depende exclusivamente de grandes deslocamentos locais para apresentar associação com o ganho global.

A Tabela 4.6 apresenta os resultados considerando a compatibilidade pessimista, definida a partir do maior valor da métrica de dissimilaridade observado entre os clientes na rodada 2. De forma semelhante ao observado para a compatibilidade média, os coeficientes de Pearson

experimento 3

	pearson	spearman
fid	-0.805	-0.100
js	-0.673	-0.100
kl_divergence	-0.653	-0.100
mmd	-0.852	-0.100
delta_global_accuracy	1.000	1.000

Tabela 4.5 – Tabela do experimento 3, que correlaciona a compatibilidade entre clientes a partir da média das métricas de transferabilidade com o delta de acurácia global do modelo

experimento 3

	pearson	spearman
fid	-0.767	-0.100
js	-0.826	-0.100
kl_divergence	-0.833	-0.100
mmd	-0.884	-0.100
delta_global_accuracy	1.000	1.000

Tabela 4.6 – Tabela do experimento 3, que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo.

permanecem negativos e de magnitude moderada a elevada, variando aproximadamente entre $-0{,}76$ e $-0{,}88$. Isso indica que, mesmo ao considerar apenas o cliente mais discrepante, ainda há associação linear relevante entre incompatibilidade inicial e ganho global ΔA. Entretanto, os coeficientes de Spearman permanecem próximos de $-0{,}1$, evidenciando ausência de relação monotônica consistente entre as variáveis. Assim, embora exista tendência linear negativa, a ordenação dos níveis de incompatibilidade não se traduz diretamente na ordenação dos ganhos globais. A similaridade entre os resultados da compatibilidade média e pessimista sugere que, neste cenário com apenas uma época local, o comportamento global da federação não é dominado por um cliente extremo específico. Em vez disso, o enfraquecimento da relação monotônica

parece estar associado à menor intensidade das atualizações locais, que reduzem o afastamento estatístico entre modelos e, consequentemente, a sensibilidade do indicador de compatibilidade.

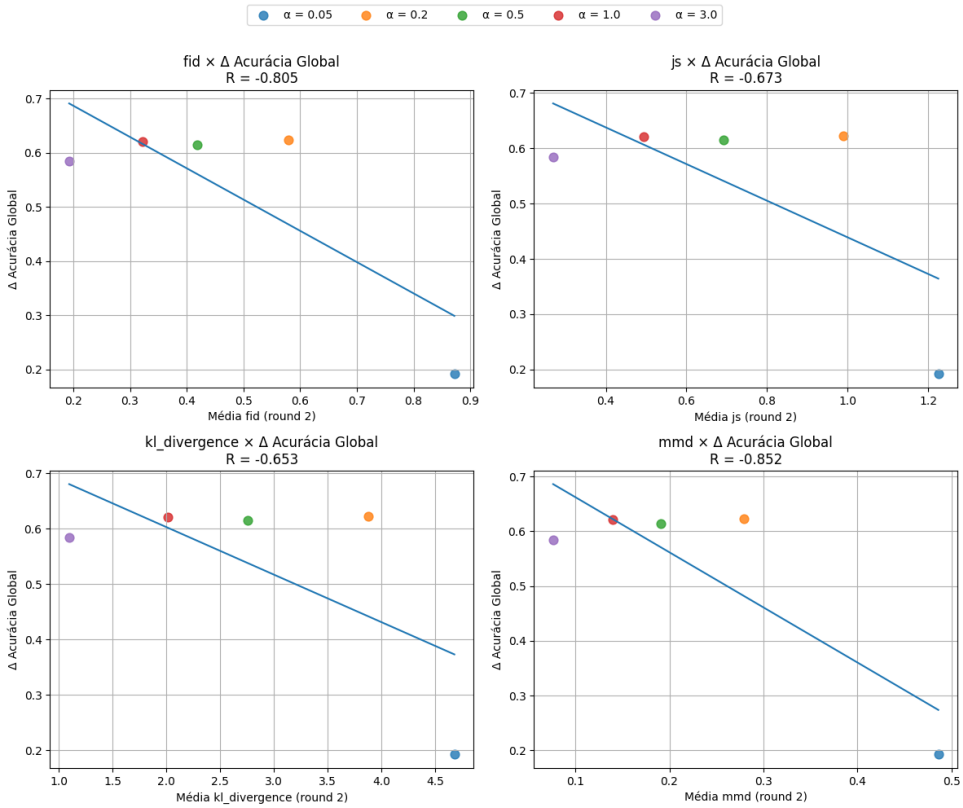

Figura 4.3 – Gráfico do experimento 3 que mostra a relação entre compatibilidade médiae o delta global de acurácia.

4.1.4 Experimento 4

No Experimento 4, o conjunto de dados permanece o CIFAR-10, porém a arquitetura da rede neural é alterada. Diferentemente dos Experimentos 1, 2 e 3, nos quais foi utilizada a ResNet-18, neste cenário emprega-se a arquitetura LightCNN, conforme descrito na Tabela 3.1.

A Tabela 4.7 apresenta as correlações entre a compatibilidade média e o ganho global ΔA. Observa-se que os coeficientes permanecem fortemente negativos para todas as métricas analisadas, com valores de Pearson próximos de $-0{,}9$ e coeficientes de Spearman em torno de $-0{,}9$, indicando relação monotônica intensa entre incompatibilidade inicial e desempenho final do modelo.

As métricas baseadas em representações intermediárias (FID e MMD) continuam apresentando correlações elevadas, sugerindo que o deslocamento no espaço de *features* permanece fortemente associado ao ganho global mesmo sob mudança arquitetural. De forma consistente, as divergências probabilísticas (KL e JS) também mantêm comportamento semelhante ao observado com a ResNet-18.

experimento 4

	pearson	spearman
fid	-0.932	-0.900
js	-0.882	-0.900
kl_divergence	-0.885	-0.900
mmd	-0.913	-0.900
delta_global_accuracy	1.000	1.000

Tabela 4.7 – Tabela do experimento 4, que correlaciona a compatibilidade média com o delta de acurácia global do modelo.

A proximidade dos resultados obtidos com diferentes arquiteturas indica que o conceito de compatibilidade entre clientes não está restrito a uma estrutura específica de rede neural. Isso sugere que a relação observada entre incompatibilidade estatística inicial e ganho global é robusta à variação do modelo utilizado.

experimento 4

	pearson	spearman
fid	-0.947	-0.900
js	-0.826	-0.900
kl_divergence	-0.828	-0.900
mmd	-0.951	-0.900
delta_global_accuracy	1.000	1.000

Tabela 4.8 – Tabela do experimento 4, que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo.

A Tabela 4.8 apresenta os resultados considerando a compatibilidade pessimista, definida a partir do maior valor da métrica de dissimilaridade observado entre os clientes na rodada 2.

Observa-se que os coeficientes de correlação permanecem fortemente negativos para todas as métricas analisadas, com valores de Pearson próximos de -1 e coeficientes de Spearman em torno de $-0{,}9$. Esse comportamento é consistente com o observado para a compatibilidade

média, indicando que o pior caso de incompatibilidade entre clientes mantém associação intensa com o ganho global ΔA. As métricas baseadas em representações intermediárias (FID e MMD) continuam apresentando os maiores coeficientes em magnitude, enquanto as divergências probabilísticas (KL e JS) mantêm padrão semelhante ao já observado nos experimentos anteriores.

A estabilidade dos resultados entre as versões média e pessimista reforça a robustez do indicador de compatibilidade mesmo sob mudança arquitetural, sugerindo que a dinâmica observada não depende da escolha específica da rede neural utilizada.

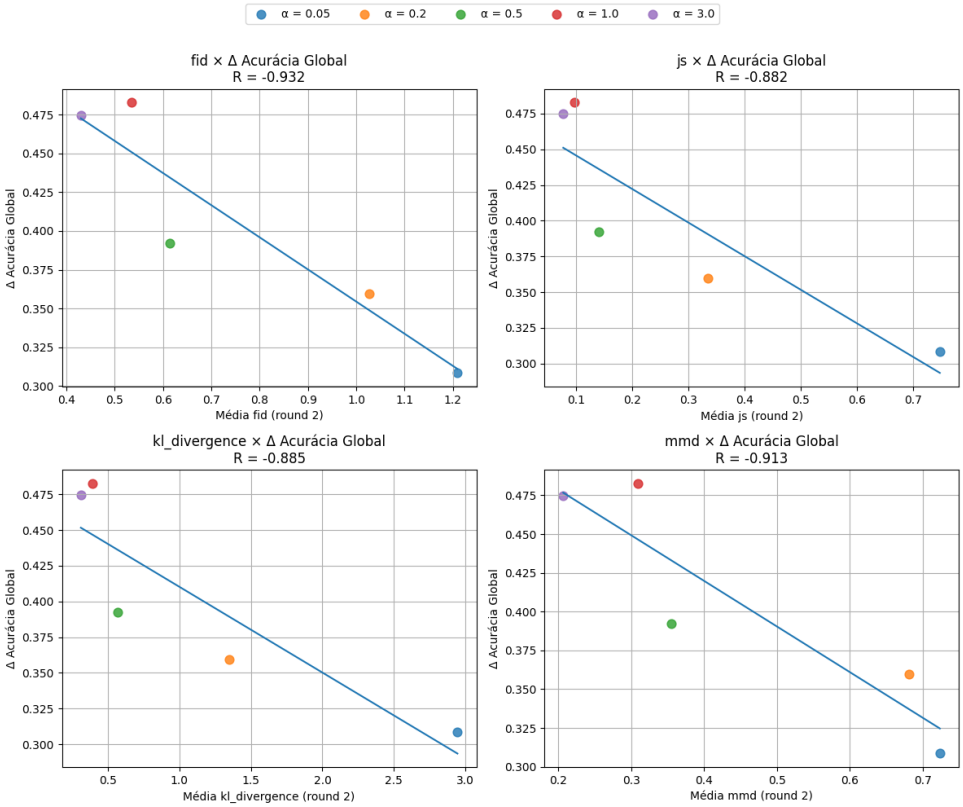

Figura 4.4 – Gráfico do experimento 4 que mostra a relação entre compatibilidade médiae o delta global de acurácia.

Um aspecto relevante deste experimento é que a LightCNN não foi inicializada com pesos pré-treinados, além de possuir arquitetura significativamente mais simples quando comparada à ResNet-18. Ainda assim, os indicadores de compatibilidade apresentaram magnitudes muito semelhantes às observadas nos experimentos anteriores.

Embora o ganho global ΔA tenha variado entre 0% e 47%, valores inferiores aos obtidos com a ResNet-18, a associação entre compatibilidade inicial e desempenho final permaneceu intensa. A redução do ΔA pode ser atribuída à menor capacidade representacional da arquitetura e à ausência de pré-treinamento, fatores que naturalmente limitam o desempenho máximo alcançável pelo modelo.

O gráfico apresentado na Figura 4.4 ilustra a relação entre a compatibilidade média e o

ganho global ΔA. Observa-se que, mesmo em um cenário com menor capacidade de modelagem, os níveis de incompatibilidade estatística nas rodadas iniciais continuam fortemente associados ao desempenho final do modelo federado.

Esse resultado sugere que o indicador de compatibilidade proposto pode estar capturando propriedades inerentes às relações estatísticas entre os clientes, isto é, ao grau de alinhamento ou discrepância entre suas distribuições locais. Dessa forma, mesmo em uma arquitetura mais simples e sem pré-treinamento, a medida de compatibilidade permanece informativa acerca da interação entre os conjuntos de dados distribuídos na federação.

Isso indica que as relações de compatibilidade observadas não dependem exclusivamente da complexidade arquitetural do modelo utilizado. Em particular, é plausível que, sob uma arquitetura com maior capacidade representacional, os mesmos padrões de relação entre clientes pudessem resultar em ganhos globais ainda mais expressivos. Investigações adicionais seriam necessárias para aprofundar a compreensão desses mecanismos.

4.1.5 Experimento 5

O Experimento 5 mantém a mesma configuração do Experimento 4, conjunto CIFAR-10, arquitetura LightCNN, 10 rodadas globais e 10 épocas locais. Diferindo apenas no número de clientes, que passa de 10 para 25, conforme descrito na Tabela 3.1.

experimento 5

	pearson	spearman
fid	-0.970	-0.900
js	-0.700	-0.900
kl_divergence	-0.700	-0.900
mmd	-0.979	-0.900
delta_global_accuracy	1.000	1.000

Tabela 4.9 – Tabela do experimento 5, que correlaciona a compatibilidade média com o delta de acurácia global do modelo.

A Tabela 4.9 apresenta as correlações entre a compatibilidade média e o ganho global ΔA. De maneira geral, observa-se que os coeficientes permanecem fortemente negativos para todas as métricas analisadas, indicando que o aumento no número de clientes não altera substancialmente o padrão observado no Experimento 4.

As métricas baseadas em representações intermediárias (FID e MMD) apresentam correlações lineares particularmente intensas, com coeficientes de Pearson próximos de -1, sugerindo relação quase linear entre incompatibilidade inicial e desempenho final.

Por outro lado, as métricas baseadas em divergências probabilísticas (KL e JS) também mantêm correlação negativa relevante, embora com magnitude ligeiramente inferior em comparação às métricas baseadas em *features*. Ainda assim, o comportamento monotônico permanece consistente, como indicado pelos coeficientes de Spearman.

De forma semelhante aos experimentos anteriores, os resultados reforçam a estabilidade do indicador de compatibilidade frente ao aumento da cardinalidade da federação, mesmo quando empregada uma arquitetura mais simples e sem pré-treinamento.

experimento 5

	pearson	spearman
fid	-0.928	-0.800
js	-0.622	-0.900
kl_divergence	-0.621	-0.900
mmd	-0.975	-0.900
delta_global_accuracy	1.000	1.000

Tabela 4.10 – Tabela do experimento 5, que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo.

A Tabela 4.10 apresenta os resultados considerando a compatibilidade pessimista, definida a partir do maior valor da métrica de dissimilaridade observado entre os clientes na rodada 2.

De modo geral, os resultados permanecem consistentes com aqueles obtidos para a compatibilidade média. As métricas baseadas em representações intermediárias (FID e MMD) continuam apresentando correlações negativas elevadas com o ganho global ΔA, com destaque para a MMD, cujo coeficiente de Pearson permanece próximo de -1.

Por outro lado, as métricas baseadas em divergências probabilísticas (KL e JS) exibem redução adicional na magnitude da correlação linear, quando comparadas à versão média. Embora os coeficientes de Spearman ainda indiquem associação monotônica negativa relevante, observa-se menor intensidade na relação linear com ΔA.

Esse comportamento sugere que, neste cenário com maior número de clientes e arquitetura simplificada, o pior caso de incompatibilidade afeta de maneira mais pronunciada as

métricas baseadas em divergências informacionais do que aquelas fundamentadas no espaço de *features*. Ainda assim, o padrão geral de associação negativa entre compatibilidade e ganho global permanece preservado.

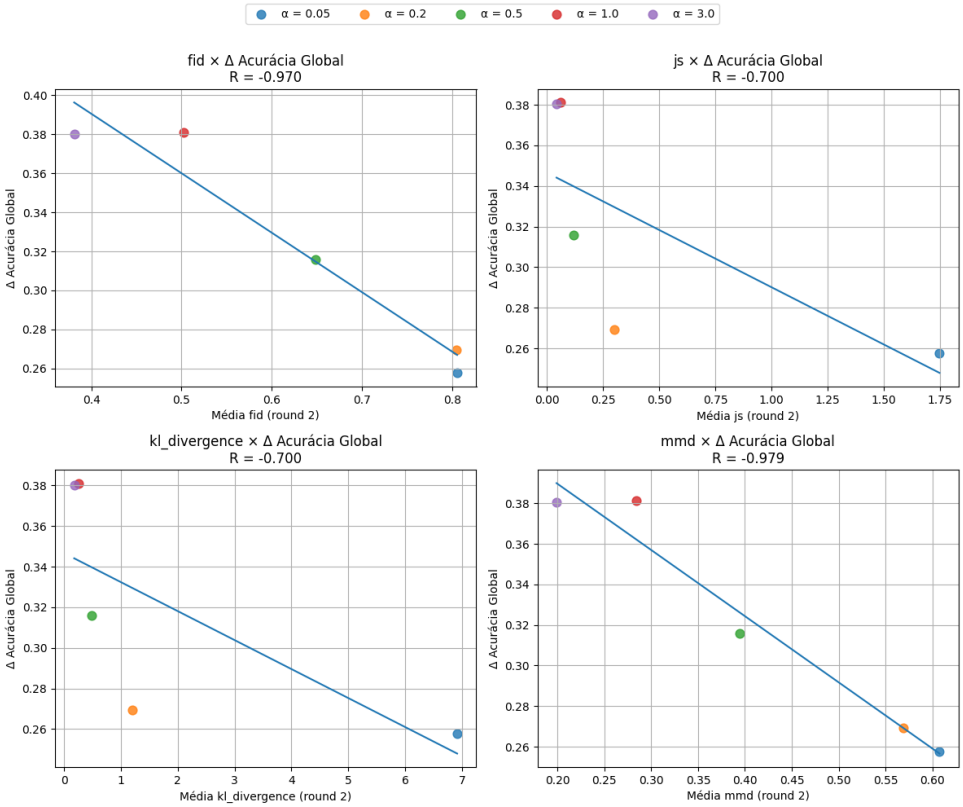

Figura 4.5 – Gráfico do experimento 5 que mostra a relação entre compatibilidade média e o delta global de acurácia.

O gráfico da figura 4.5 também mostra o relacionamento entre compatibilidade média e delta de acurácia global, e mostra também que o modelo global não obteve uma grande melhora geral de forma similar ao 4.7.

4.2 Experimentos Fashion Mnist

4.2.1 Experimento 6

O Experimento 6 introduz uma mudança no conjunto de dados, substituindo o CIFAR-10 pelo Fashion-MNIST. A arquitetura utilizada retorna à ResNet-18, conforme descrito na Tabela 3.2, mantendo-se 10 clientes, 10 rodadas globais e 10 épocas locais.

A Tabela 4.11 apresenta as correlações entre a compatibilidade média e o ganho global ΔA. Observa-se que todas as métricas analisadas mantêm correlação negativa elevada com ΔA, com coeficientes de Pearson variando aproximadamente entre $-0{,}8$ e $-0{,}9$ e coeficientes de Spearman em torno de $-0{,}8$.

experimento 6

	pearson	spearman
fid	-0.879	-0.800
js	-0.812	-0.800
kl_divergence	-0.817	-0.800
mmd	-0.908	-0.800
delta_global_accuracy	1.000	1.000

Tabela 4.11 – Tabela do experimento 6, que correlaciona a compatibilidade média com o delta de acurácia global do modelo.

As métricas baseadas em representações intermediárias (FID e MMD) continuam apresentando as maiores magnitudes de correlação linear, enquanto as divergências probabilísticas (KL e JS) exibem comportamento semelhante, ainda que ligeiramente inferior em magnitude.

Esses resultados indicam que o conceito de compatibilidade entre clientes mantém associação consistente com o ganho global mesmo sob mudança de *dataset*. A estabilidade do padrão observado sugere que a relação entre incompatibilidade estatística inicial e desempenho final não é restrita a um conjunto de dados específico, reforçando o potencial de generalização do indicador proposto.

experimento 6

	pearson	spearman
fid	-0.934	-0.900
js	-0.947	-0.800
kl_divergence	-0.958	-0.800
mmd	-0.876	-0.800
delta_global_accuracy	1.000	1.000

Tabela 4.12 – Tabela de correlação entre compatibilide pessimista e delta global.

A Tabela 4.12 apresenta os resultados considerando a compatibilidade pessimista. De

modo geral, não se observam alterações substanciais em relação à versão média.

As correlações permanecem fortemente negativas para todas as métricas analisadas, com magnitudes similares às anteriormente reportadas. Pequenas variações nos coeficientes não alteram o padrão geral de associação entre incompatibilidade inicial e ganho global ΔA.

Esse comportamento indica que, no cenário do Fashion-MNIST com 10 clientes, o pior caso de incompatibilidade não exerce influência qualitativamente distinta daquela observada na compatibilidade média, preservando o mesmo padrão de relação com o desempenho final do modelo global.

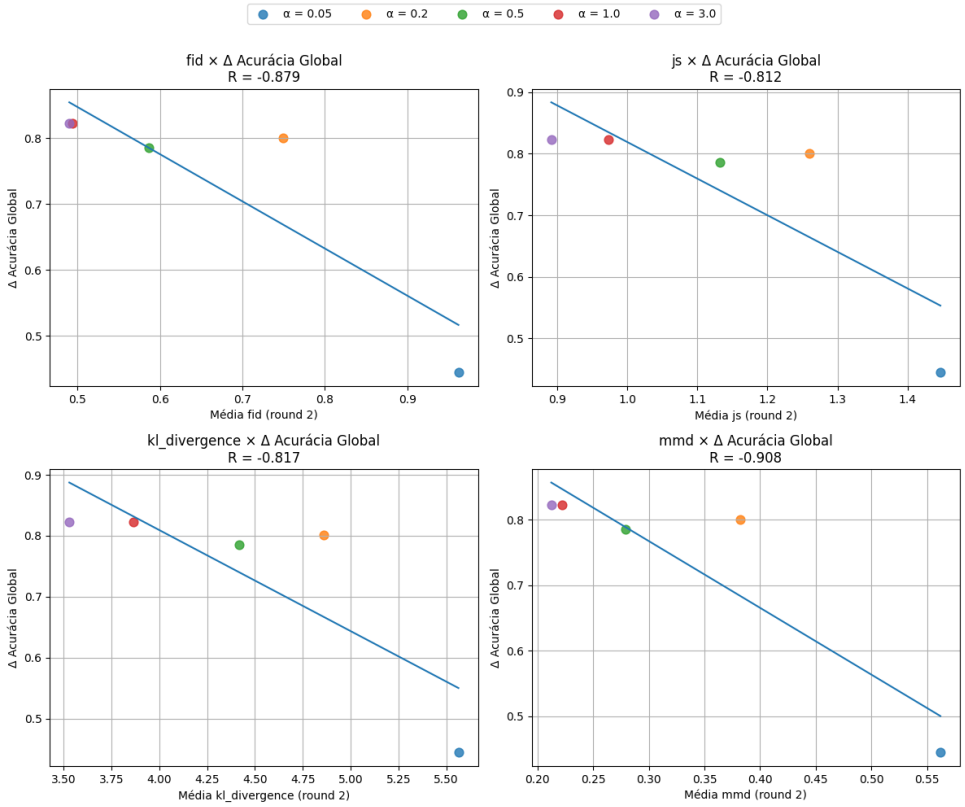

Figura 4.6 – Gráfico do experimento 6 que mostra a relação entre compatibilidade média e o delta global de acurácia.

A Figura 4.6 apresenta a relação entre a compatibilidade média e o ganho global ΔA. Diferentemente dos experimentos conduzidos com a arquitetura LightCNN, observa-se neste caso a presença de valores mais elevados de ΔA.

Esse comportamento sugere que os ganhos globais reduzidos observados nos experimentos anteriores estavam associados à menor capacidade representacional da LightCNN para a tarefa em questão, e não a uma limitação intrínseca do indicador de compatibilidade.

Com a utilização da ResNet-18 no conjunto Fashion-MNIST, o modelo volta a apresentar ganhos mais expressivos, enquanto a relação entre incompatibilidade inicial e desempenho final permanece consistente. Esse resultado reforça a interpretação de que o indicador de compatibili-

dade reflete predominantemente as relações estatísticas entre os clientes, sendo relativamente independente da capacidade absoluta do modelo utilizado.

4.2.2 Experimento 7

O Experimento 7 mantém configuração semelhante ao Experimento 6. Conjunto Fashion-MNIST, arquitetura ResNet-18, 10 rodadas globais e 10 épocas locais. Diferindo apenas no número de clientes, que passa de 10 para 25, conforme apresentado na Tabela 3.2.

experimento 7

	pearson	spearman
fid	-0.686	-0.600
js	-0.457	-0.500
kl_divergence	-0.394	-0.500
mmd	-0.778	-0.700
delta_global_accuracy	1.000	1.000

Tabela 4.13 – Tabela do experimento 7, que correlaciona a compatibilidade média com o delta de acurácia global do modelo.

A Tabela 4.13 evidencia redução na magnitude dos coeficientes de correlação quando comparados ao Experimento 6. Observa-se queda mais pronunciada nas métricas baseadas em divergências probabilísticas (KL e JS), enquanto as métricas baseadas em representações intermediárias (FID e MMD) mantêm associação negativa relevante, ainda que inferior à observada anteriormente.

Esse comportamento indica enfraquecimento parcial da relação entre compatibilidade inicial e ganho global ΔA. Entretanto, não é possível concluir que houve deterioração estrutural do indicador de compatibilidade, uma vez que experimentos análogos, como o Experimento 2, também conduzido com 25 clientes, também apresentaram correlações elevadas.

Uma possível explicação para essa variação reside na realização específica da distribuição de dados via particionamento Dirichlet. Como a geração das partições depende de inicialização aleatória, diferentes sementes podem produzir níveis distintos de heterogeneidade efetiva, mesmo mantendo fixo o parâmetro α. Dessa forma, o enfraquecimento observado pode estar associado a uma configuração particular de distribuição entre clientes, e não necessariamente a uma limitação conceitual do indicador proposto.

experimento 7

	pearson	spearman
fid	-0.801	-0.700
js	-0.689	-0.900
kl_divergence	-0.688	-0.900
mmd	-0.768	-0.600
delta_global_accuracy	1.000	1.000

Tabela 4.14 – Tabela do experimento 7, que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo.

A Tabela 4.14 apresenta os resultados considerando a compatibilidade pessimista. Diferentemente do observado na compatibilidade média, verifica-se aumento na magnitude dos coeficientes de correlação para a maioria das métricas.

Em particular, as métricas baseadas em representações intermediárias (FID e MMD) voltam a apresentar associação negativa mais intensa com o ganho global ΔA, quando comparadas à versão média. As divergências probabilísticas (KL e JS) também mantêm correlação negativa relevante.

Esse comportamento sugere que, neste experimento específico, o desempenho global pode estar mais fortemente associado ao cliente mais discrepante do que ao comportamento agregado da federação. Em outras palavras, a heterogeneidade extrema capturada pela compatibilidade pessimista, parece exercer influência mais pronunciada do que a incompatibilidade média entre os clientes.

Dessa forma, o enfraquecimento observado na compatibilidade média pode estar relacionado à diluição do efeito de clientes extremos quando se utiliza agregação por média, enquanto a abordagem pessimista preserva a influência desses casos mais críticos.

O gráfico da figura 4.7 indica que o modelo conseguiu um aprendizado alto a partir dos indices de delta de acurácia global, mesmo com indices de compatibilidade não tão altos.

4.2.3 Experimento 8

O Experimento 8 amplia significativamente a escala da federação, sendo conduzido com 100 clientes e 20 rodadas globais, conforme descrito na Tabela 3.2. O objetivo principal foi avaliar o comportamento do indicador de compatibilidade em cenários com elevado número de

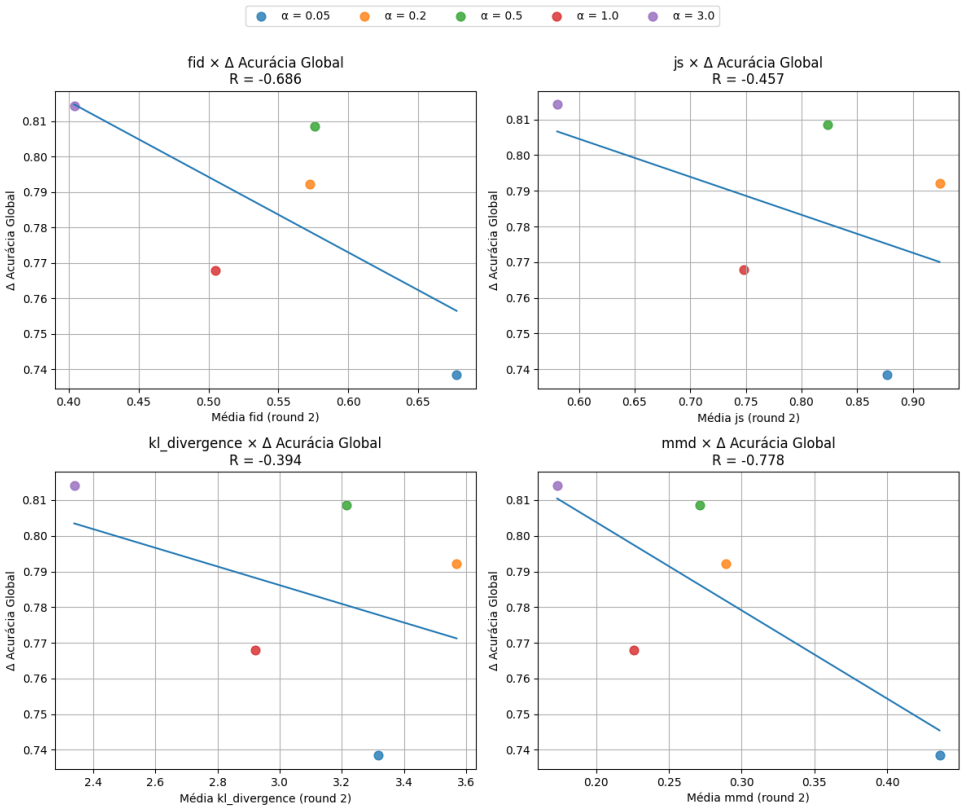

Figura 4.7 – Gráfico do experimento 7 que mostra a relação entre compatibilidade média e o delta global de acurácia.

clientes, nos quais a heterogeneidade tende a ser mais acentuada e o balanceamento de classes se torna mais desafiador.

Durante a implementação deste experimento, observou-se uma limitação prática relacionada à alocação mínima de classes por cliente. Devido à combinação entre número elevado de clientes e número reduzido de classes do conjunto Fashion-MNIST, foi necessário reduzir a restrição mínima de classes por cliente para viabilizar o particionamento via distribuição Dirichlet. Essa adaptação aumentou potencialmente o grau de desbalanceamento local em comparação aos experimentos anteriores.

A Tabela 4.15 apresenta as correlações entre a compatibilidade média e o ganho global ΔA. Apesar da maior complexidade do cenário, todas as métricas mantêm correlação negativa elevada, com coeficientes de Pearson próximos de $-0{,}9$ e coeficientes de Spearman variando entre $-0{,}8$ e $-0{,}9$.

Esse resultado indica que, mesmo sob condições mais extremas de heterogeneidade e maior número de clientes, o indicador de compatibilidade preserva associação consistente com o desempenho final do modelo global. Tal comportamento reforça a robustez da medida proposta em cenários de larga escala.

A Tabela 4.16 apresenta os resultados considerando a compatibilidade pessimista. De

experimento 8

	pearson	spearman
fid	-0.902	-0.900
js	-0.923	-0.800
kl_divergence	-0.904	-0.800
mmd	-0.917	-0.900
delta_global_accuracy	1.000	1.000

Tabela 4.15 – Tabela do experimento 8, que correlaciona a compatibilidade média com o delta de acurácia global do modelo.

experimento 8

	pearson	spearman
fid	-0.955	-0.700
js	-0.747	-1.000
kl_divergence	-0.735	-1.000
mmd	-0.980	-0.600
delta_global_accuracy	1.000	1.000

Tabela 4.16 – Tabela do experimento 8, que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo.

maneira geral, os coeficientes permanecem fortemente negativos, preservando o padrão de associação entre incompatibilidade inicial e ganho global ΔA mesmo no cenário com 100 clientes.

As métricas baseadas em representações intermediárias (FID e MMD) continuam apresentando correlações lineares elevadas, enquanto as divergências probabilísticas (KL e JS) exibem maior variação entre correlação linear e monotônica. Ainda assim, o padrão geral de relação negativa permanece consistente.

Esses resultados reforçam que, mesmo em um cenário de maior escala e heterogeneidade acentuada, o pior caso de incompatibilidade mantém associação relevante com o desempenho

final do modelo federado.

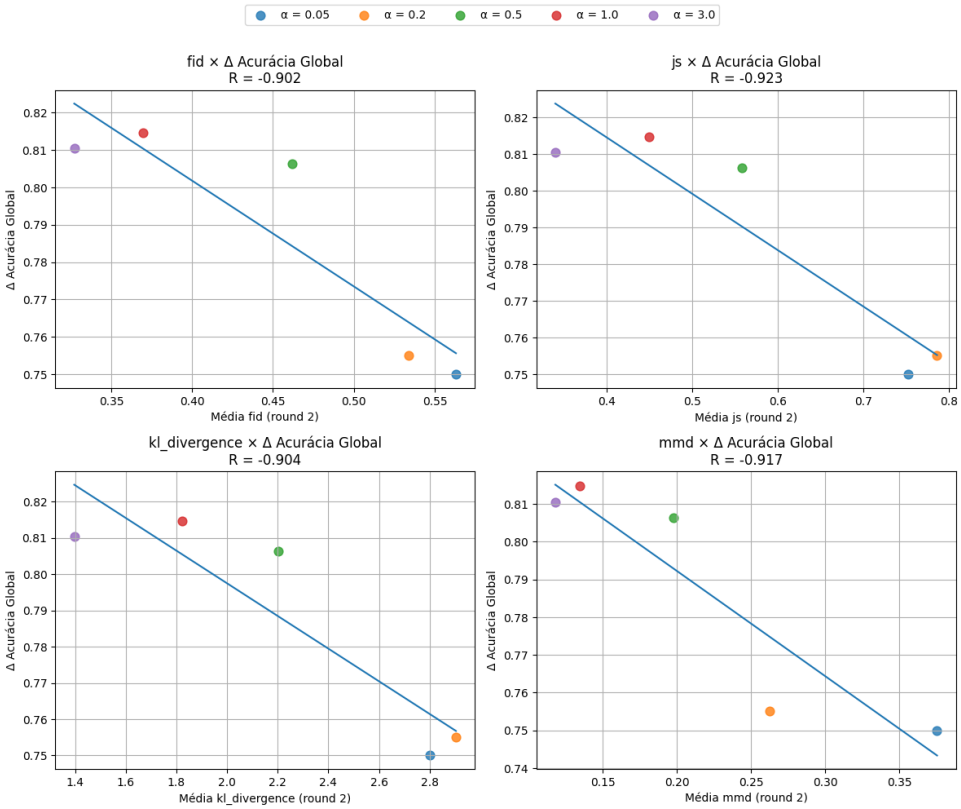

Figura 4.8 – Gráfico do experimento 8 que mostra a relação entre compatibilidade média e o delta global de acurácia.

O gráfico do experimento 8 pode ser analisado na figura 4.8, e reflete que o modelo apresentou uma boa taxa de aprendizado a partir do delta de acurácia global.

4.3 Experimentos Blood Mnist

Nos experimentos a seguir, altera-se novamente o conjunto de dados, mantendo-se o número reduzido de épocas locais e repetindo cada configuração 30 vezes, com o objetivo de obter maior robustez estatística dos resultados.

A escolha por um número baixo de épocas locais decorre do fato de que a análise de compatibilidade é realizada entre o modelo global da rodada 1 e o modelo local da rodada 2. Atualizações locais excessivamente longas poderiam produzir afastamentos muito intensos entre os modelos, ampliando artificialmente as discrepâncias estatísticas capturadas pelas métricas. Ao limitar o número de épocas locais, busca-se preservar um regime de atualização mais moderado, permitindo avaliar a sensibilidade do indicador de compatibilidade em cenários com deslocamentos menos acentuados.

Além disso, configurações semelhantes, com número reduzido de épocas locais, têm sido adotadas em estudos envolvendo conjuntos de dados médicos em cenários não-IID, devido à maior variabilidade e sensibilidade dessas bases.

4.3.1 Experimento 9

O Experimento 9 foi conduzido utilizando o conjunto Blood-MNIST, arquitetura ResNet-18 e 2 épocas locais, conforme descrito na Tabela 3.3. Diferentemente dos experimentos anteriores, cada configuração foi repetida 30 vezes, visando maior robustez estatística.

experimento 9

	pearson	spearman
fid	0.171	0.141
js	0.770	0.775
kl_divergence	0.766	0.774
mmd	-0.827	-0.799
delta_global_accuracy	1.000	1.000

Tabela 4.17 – Tabela do experimento 9, que correlaciona a compatibilidade média com o delta de acurácia global do modelo.

A Tabela 4.17 revela um comportamento distinto daquele observado nos conjuntos CIFAR-10 e Fashion-MNIST. A métrica MMD mantém correlação negativa elevada com o ganho global ΔA, em consonância com a hipótese central de que maior discrepância no espaço de representações internas está associada a menor desempenho final do modelo global.

Por outro lado, a FID apresenta correlação fraca, indicando menor sensibilidade à dinâmica observada neste conjunto de dados.

As métricas baseadas em divergências probabilísticas (KL e JS) exibem correlação positiva moderada com ΔA. Esse comportamento contrasta com os resultados obtidos nos experimentos anteriores, nos quais tais métricas apresentaram associação negativa. No cenário do Blood-MNIST, maiores valores iniciais de divergência probabilística entre modelo global e modelos locais estão associados a maiores ganhos globais.

Esse resultado sugere que, neste conjunto de dados, a divergência nas distribuições de saída(*logits*) pode refletir maior diversidade informativa entre clientes, potencialmente contribuindo para o processo de agregação. No entanto, como esse padrão não foi observado nos

demais conjuntos de dados, sua interpretação deve ser feita com cautela, indicando que a relação entre divergência probabilística inicial e ganho global pode ser dependente do contexto e das características específicas do dataset analisado.

experimento 9

	pearson	spearman
fid	-0.042	-0.050
js	0.405	0.487
kl_divergence	0.361	0.441
mmd	-0.834	-0.800
delta_global_accuracy	1.000	1.000

Tabela 4.18 – Tabela do experimento 9, que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo.

A Tabela 4.18 apresenta os resultados considerando a compatibilidade pessimista. Diferentemente do observado na versão média, as métricas baseadas em divergências probabilísticas (KL e JS) passam a apresentar correlações positivas de baixa magnitude, com coeficientes próximos de $0{,}4$.

Esse enfraquecimento indica que a associação positiva observada anteriormente não se mantém de forma robusta quando se considera o pior caso de incompatibilidade entre clientes. Dessa forma, não é possível afirmar que a divergência probabilística inicial possua relação consistente com o ganho global ΔA neste cenário.

A FID, por sua vez, mantém correlação próxima de zero, reforçando sua sensibilidade ao conjunto Blood-MNIST e sua limitada capacidade explicativa neste experimento.

Em contraste, a MMD preserva correlação negativa elevada tanto na versão média quanto na pessimista, mantendo coerência com a hipótese central do trabalho. Esse comportamento sugere que a discrepância no espaço de representações internas se mostra mais estável e consistente como indicador de compatibilidade no contexto analisado.

A figura 4.9 mostra os gráficos de relação da compatibilidade média com o delta de acurácia global. Para esse caso é possível notar a grande quantidade de pontos gerados pelos diferentes experimentos.

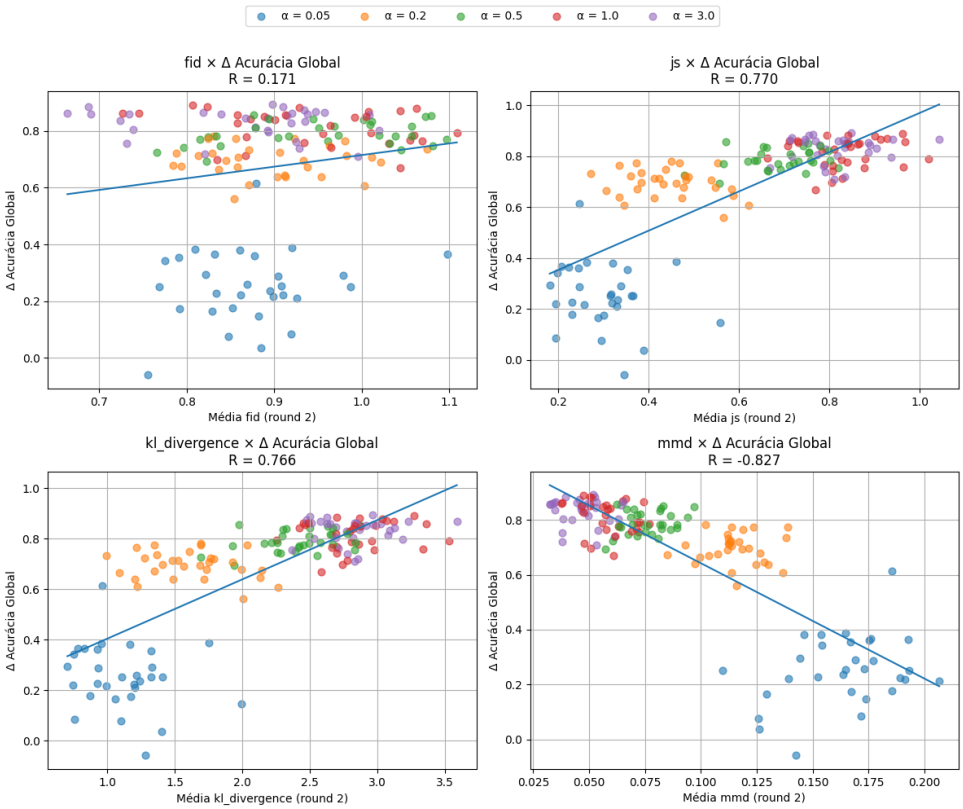

Figura 4.9 – Gráfico do experimento 9 que mostra a relação entre compatibilidade média e o delta global de acurácia.

4.3.2 Experimento 10

O Experimento 10 mantém a mesma configuração do Experimento 9. Conjunto Blood-MNIST, arquitetura ResNet-18 e 2 épocas locais, diferindo apenas no número de clientes, que passa de 10 para 25, conforme descrito na Tabela 3.3. Assim como no experimento anterior, cada configuração foi repetida 30 vezes.

A Tabela 4.19 apresenta as correlações entre a compatibilidade média e o ganho global ΔA. Assim como observado no Experimento 9, o comportamento das métricas no conjunto Blood-MNIST difere daquele verificado nos experimentos com CIFAR-10 e Fashion-MNIST.

A métrica MMD mantém correlação negativa elevada, preservando coerência com a hipótese de que maior discrepância no espaço de representações internas está associada a menor desempenho final do modelo global.

Em contraste, as métricas baseadas em divergências probabilísticas (KL e JS) apresentam correlação positiva forte com ΔA, indicando padrão distinto do observado nos demais conjuntos de dados. A FID também passa a exibir correlação positiva relevante neste cenário.

Embora esse comportamento se repita em relação ao experimento anterior com Blood-MNIST, sua interpretação deve ser realizada com cautela, uma vez que tal padrão não foi identifi-

experimento 10

	pearson	spearman
fid	0.727	0.640
js	0.827	0.775
kl_divergence	0.811	0.757
mmd	-0.849	-0.811
delta_global_accuracy	1.000	1.000

Tabela 4.19 – Tabela do experimento 10, que correlaciona a compatibilidade média com o delta de acurácia global do modelo.

cado nos demais datasets analisados. Isso sugere que a associação entre divergência probabilística inicial e ganho global pode depender das características específicas do conjunto de dados e do regime de agregação considerado.

De forma geral, a MMD permanece como o indicador mais estável no cenário analisado, ao manter coerência com a hipótese central mesmo sob variações no número de clientes.

experimento 10 max

	pearson	spearman
fid	0.068	0.083
js	0.154	0.272
kl_divergence	0.106	0.234
mmd	-0.798	-0.770
delta_global_accuracy	1.000	1.000

Figura 4.10 – Tabela do experimento 10, que correlaciona a compatibilidade pessimista com o delta de acurácia global do modelo.

A Tabela 4.10 apresenta os resultados considerando a compatibilidade pessimista. Observa-se que, diferentemente da versão média, as métricas baseadas em divergências probabilísticas (KL e JS) passam a apresentar correlações positivas de baixa magnitude, enquanto a FID exibe coeficientes próximos de zero.

Esse enfraquecimento indica que a associação positiva observada na compatibilidade média não se mantém de forma robusta quando se considera o pior caso de incompatibilidade entre clientes. Assim, não é possível afirmar que a divergência probabilística inicial possua relação consistente com o ganho global ΔA neste cenário.

Em contraste, a MMD preserva correlação negativa elevada tanto na versão média quanto na pessimista, mantendo alinhamento com a hipótese central do trabalho. Esse comportamento reforça a estabilidade da MMD como indicador de compatibilidade no contexto do Blood-MNIST, mesmo sob variações no número de clientes e no critério de agregação.

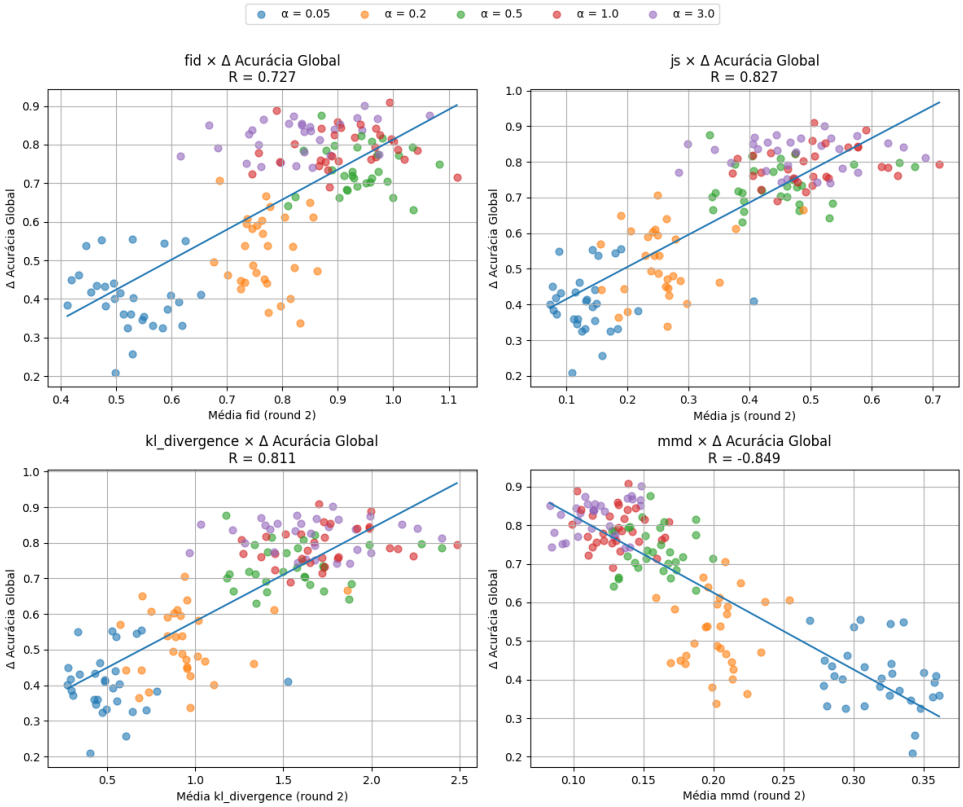

Figura 4.11 – Gráfico do experimento 10 que mostra a relação entre compatibilidade média e o delta global de acurácia.

A Figura 4.11 apresenta a relação entre a compatibilidade média — calculada a partir de diferentes métricas de transferabilidade no segundo round, e o ganho global de acurácia ΔA para todas as execuções dos experimentos com o conjunto Blood-MNIST.

5 Considerações Finais

5.1 Conclusão

Este trabalho investigou o conceito de compatibilidade entre clientes em cenários de aprendizado federado, definindo-a como a discrepância estatística entre o modelo global da rodada anterior e os modelos locais atualizados, avaliada sobre os dados de cada cliente. A hipótese central consistiu em analisar se tal medida, calculada nas rodadas iniciais do treinamento, estaria associada ao ganho global de desempenho ΔA ao final do processo federado.

Os resultados obtidos ao longo dos diferentes cenários experimentais permitem destacar alguns pontos centrais.

Primeiramente, a métrica baseada em representações intermediárias MMD apresentou comportamento consistente em todos os conjuntos de dados analisados (CIFAR-10, Fashion-MNIST e Blood-MNIST), mantendo correlação negativa elevada com o ganho global tanto na versão média quanto na pessimista. Esse padrão indica que maior discrepância no espaço de *features* entre modelos locais e global tende a estar associada a menor desempenho final da federação, de forma robusta a variações de arquitetura, número de clientes e escala do experimento.

Em contraste, a FID demonstrou instabilidade, especialmente nos experimentos com o Blood-MNIST, nos quais perdeu poder explicativo. Tal comportamento sugere sensibilidade dessa métrica às características específicas do conjunto de dados.

As métricas baseadas em divergências probabilísticas (KL e JS) apresentaram comportamento dependente do contexto. Enquanto nos conjuntos CIFAR-10 e Fashion-MNIST exibiram correlação negativa consistente com ΔA, nos experimentos com Blood-MNIST passaram a apresentar correlação positiva na versão média e correlação enfraquecida na versão pessimista. Esse resultado indica que tais métricas não se mostraram estáveis sob diferentes regimes de agregação e características de dados, sugerindo que sua interpretação requer cautela.

De maneira geral, os experimentos evidenciam que nem todas as medidas de divergência estatística capturam de forma igualmente robusta a dinâmica de compatibilidade entre clientes. Entre as métricas analisadas, a MMD destacou-se como o indicador mais estável e consistente ao longo dos diferentes cenários.

Além da análise retrospectiva dos experimentos, um aspecto particularmente relevante deste trabalho é o potencial uso preditivo da compatibilidade baseada em MMD. Como a medida é calculada a partir das duas primeiras rodadas do treinamento federado, antes da convergência do modelo global, ela pode funcionar como um indicador antecipado do desempenho final da federação.

Em um cenário prático, considere uma federação recém-inicializada em que, após a segunda rodada, calcula-se a compatibilidade média via MMD entre modelos locais e o modelo global. Caso o valor obtido seja elevado (indicando grande discrepância no espaço de representações), os resultados apresentados neste trabalho sugerem que o experimento tende a apresentar ganho global reduzido ao final do treinamento. Por outro lado, valores baixos de MMD estariam associados a maior probabilidade de bom desempenho global.

Essa propriedade abre possibilidade para aplicações como:

- Diagnóstico precoce de federações potencialmente instáveis;
- Ajuste dinâmico de hiperparâmetros (por exemplo, número de épocas locais ou taxa de aprendizado);
- Reagrupamento ou filtragem de clientes altamente incompatíveis antes da continuidade do treinamento.

Dessa forma, a compatibilidade baseada em MMD não apenas se mostrou estatisticamente associada ao desempenho final, mas também apresenta potencial utilidade prática como ferramenta de monitoramento e tomada de decisão em ambientes federados.

5.2 Trabalhos Futuros

Os resultados obtidos neste trabalho abrem diversas possibilidades de investigação futura, tanto no âmbito experimental quanto teórico.

Uma primeira direção consiste na ampliação dos experimentos para um conjunto mais diverso de *datasets*, especialmente em cenários com heterogeneidade acentuada (altamente non-IID). Em particular, seria relevante explorar sistematicamente configurações com número reduzido de épocas locais, uma vez que esse regime é frequentemente adotado em cenários federados com elevada heterogeneidade ou restrições computacionais. A análise detalhada da compatibilidade sob diferentes intensidades de atualização local pode contribuir para compreender melhor sua sensibilidade ao deslocamento entre modelos.

Do ponto de vista teórico, um ponto que merece investigação aprofundada é a diferença de comportamento observada entre FID e MMD, apesar de ambas operarem no espaço de representações intermediárias. Enquanto a MMD apresentou estabilidade consistente ao longo dos experimentos, a FID mostrou sensibilidade significativa, especialmente no conjunto BloodMNIST.

Outra linha promissora envolve a exploração mais exaustiva do potencial da MMD como indicador de compatibilidade. Estudos futuros poderiam avaliar:

- A sensibilidade da MMD à escolha do kernel e ao parâmetro de largura;
- O impacto da normalização das representações;
- A utilização de versões ponderadas da MMD que considerem o tamanho relativo dos clientes.

Adicionalmente, seria particularmente relevante investigar o comportamento da compatibilidade baseada em MMD em cenários adversariais. A introdução controlada de clientes maliciosos ou com dados propositalmente enviesados permitiria avaliar se a medida é capaz de sinalizar comportamentos anômalos nas rodadas iniciais, abrindo possibilidade para sua utilização como ferramenta de detecção precoce de ataques em aprendizado federado.

Por fim, investigações futuras podem explorar a integração da compatibilidade como mecanismo ativo no algoritmo federado, por exemplo:

- Ajustando dinamicamente pesos de agregação;
- Aplicando filtragem ou reagrupamento de clientes;
- Interrompendo precocemente experimentos com alta incompatibilidade inicial.

Tais extensões podem transformar a compatibilidade entre clientes de um indicador analítico para um componente operacional do treinamento federado.

Referências

3Blue1Brown. *Gradient Descent – 3Blue1Brown*. <https://www.3blue1brown.com/lessons/gradient-descent>. Disponível em: <https://www.3blue1brown.com/lessons/gradient-descent>. Acesso em: 14 jul. 2025.

BANABILAH, S.; ALOQAILY, M.; ALSAYED, E.; MALIK, N.; JARARWEH, Y. Federated learning review: Fundamentals, enabling technologies, and future applications. *Journal of Network and Computer Applications*, Elsevier, v. 215, p. 103609, 2023. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1084804523001201>. Acesso em: 18 ago. 2025.

BAO, G.; GUO, P. Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges. *Cluster Computing*, Springer, v. 24, p. 2205–2222, 2021. Disponível em: <https://doi.org/10.1007/s10586-021-03294-7>. Acesso em: 18 ago. 2025.

BAO, Y.; LI, Y.; HUANG, S.-L.; ZHANG, L.; ZHENG, L.; ZAMIR, A. R.; GUIBAS, L. J. An information-theoretic approach to transferability in task transfer learning. *arXiv*, abs/2212.10082, 2022. Disponível em: <https://arxiv.org/abs/2212.10082>. Acesso em: 18 ago. 2025.

BEUTEL, D. J.; TOPAL, T.; MATHUR, A.; QIU, X.; FERNANDEZ-MARQUES, J.; GAO, Y.; SANI, L.; LI, K. H.; PARCOLLET, T.; GUSMãO, P. P. B. d.; LANE, N. D. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020. Disponível em: <https://arxiv.org/abs/2007.14390>. Acesso em: 18 ago. 2025.

COSTA, E. J. X. Inteligência artificial aplicada à zootecnia. *Revista Brasileira de Zootecnia*, SciELO Brasil, v. 38, p. 390–396, 2009. Disponível em: <https://www.scielo.br>. Acesso em: 18 ago. 2025.

DUBEY, S. R.; SINGH, S. K.; CHAUDHURI, B. B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 2021. Accepted for publication; copyright will be transferred to Elsevier.

FAMá, F.; KALALAS, C.; LAGEN, S.; DINI, P. Measuring data similarity for efficient federated learning: A feasibility study. In: IEEE. *Proceedings of the IEEE International Conference on Communications (ICC)*. Montreal, QC, Canada: IEEE, 2021. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/9443496>. Acesso em: 18 ago. 2025.

GEORGEVICI, A. I.; TERBLANCHE, M. Neural networks and deep learning: a brief introduction. *Intensive Care Medicine*, v. 45, n. 5, p. 712–714, may 2019. Disponível em: <https://doi.org/10.1007/s00134-019-05537-w>. Acesso em: 18 ago. 2025.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial networks. *Advances in Neural Information Processing Systems*, v. 27, 2014.

GRETTON, A.; BORGWARDT, K. M.; RASCH, M. J.; SCHÖLKOPF, B.; SMOLA, A. J. A kernel two-sample test. *Journal of Machine Learning Research*, v. 13, p. 723–773, 2012.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778. Disponível em: <https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html>. Acesso em: 18 ago. 2025.

HEUSEL, M.; RAMSAUER, H.; UNTERTHINER, T.; NESSLER, B.; HOCHREITER, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2017.

ISLAM, M.; CHEN, G.; JIN, S. An overview of neural network. *American Journal of Neural Networks and Applications*, v. 5, n. 1, p. 7–11, 2019. Disponível em: <https://doi.org/10.11648/j.ajnna.20190501.12>. Acesso em: 18 ago. 2025.

IYER, V. N. *A Review on Different Techniques Used to Combat the Non-IID and Heterogeneous Nature of Data in Federated Learning: A Preprint*. 2025. Preprint, Department of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. Email: venkataramannatarajan2001@gmail.com.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. ArXiv:1412.6980.

KRIZHEVSKY, A. *Learning multiple layers of features from tiny images*. [S.l.], 2009. Disponível em: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. Acesso em: 18 ago. 2025.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951.

LEVY, A.; SHALOM, B. R.; CHALAMISH, M. A guide to similarity measures. *arXiv preprint*, 2024. Disponível em: <https://arxiv.org/abs/2408.07706>. Acesso em: 18 ago. 2025.

LI, L.; DOROSLOVACKI, M.; LOEW, M. H. Approximating the gradient of cross-entropy loss function. *IEEE Access*, v. 8, p. 111626–111635, 2020. Disponível em: <https://ieeexplore.ieee.org/document/9120227>. Acesso em: 18 ago. 2025.

LI, Z.; LIU, F.; YANG, W.; PENG, S.; ZHOU, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, v. 33, p. 6999–7019, 2021. Disponível em: <https://ieeexplore.ieee.org/document/9443364>. Acesso em: 18 ago. 2025.

LIN, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, v. 37, n. 1, p. 145–151, 1991.

LIU, W.; WANG, Z.; LIU, X.; ZENG, N.; LIU, Y.; ALSAADI, F. E. A survey of deep neural network architectures and their applications. *Neurocomputing*, v. 234, p. 11–26, 2017. Disponível em: <https://doi.org/10.1016/j.neucom.2016.12.038>. Acesso em: 18 ago. 2025.

LYU, L.; YU, H.; YANG, Q. *Threats to Federated Learning: A Survey*. 2020. ArXiv preprint. ArXiv:2003.02133. Disponível em: <https://arxiv.org/abs/2003.02133>. Acesso em: 18 ago. 2025.

MCMAHAN, B.; THAKURTA, A. *Federated Learning with Formal Differential Privacy Guarantees*. 2022. Google Research Blog. Disponível em: <https://research.google/blog/federated-learning-with-formal-differential-privacy-guarantees/>. Acesso em: 18 ago. 2025.

MCMAHAN, H. B.; MOORE, E.; RAMAGE, D.; HAMPSON, S.; ARCAS, B. Agüera y. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, p. 1–12, 2017. Disponível em: <https://arxiv.org/abs/1602.05629>. Acesso em: 18 ago. 2025.

MONTESINOS-LóPEZ, O. A.; MONTESINOS-LóPEZ, A.; CROSSA, J. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. [S.l.]: Springer Cham, 2022. Disponível em: <https://doi.org/10.1007/978-3-030-89010-0>. Acesso em: 18 ago. 2025. ISBN 978-3-030-89010-0.

NIELSEN, F. Two tales for a geometric jensen–shannon divergence. *arXiv preprint arXiv:2508.05066*, 2025. Disponível em: <https://arxiv.org/abs/2508.05066>.

NIELSEN, M. A. *Neural Networks and Deep Learning*. [S.l.]: Determination Press, 2015. Disponível em: <https://nnfs.io>. Acesso em: 18 ago. 2025.

NODA, K.; YAMAGUCHI, Y.; NAKADAI, K.; OKUNO, H. G.; OGATA, T. Audio-visual speech recognition using deep learning. *Applied Intelligence*, v. 42, n. 4, p. 722–737, 2015. Disponível em: <https://doi.org/10.1007/s10489-014-0629-7>. Acesso em: 18 ago. 2025.

O'SHEA, K.; NASH, R. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015. Disponível em: <http://arxiv.org/abs/1511.08458>. Acesso em: 18 ago. 2025.

PASZKE, A.; GROSS, S.; MASSA, F.; LERER, A.; BRADBURY, J.; CHANAN, G.; KILLEEN, T.; LIN, Z.; GIMELSHEIN, N.; ANTIGA, L.; OUTROS. Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2019. v. 32, p. 8026–8037. Disponível em: <https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>. Acesso em: 18 ago. 2025.

POUSHTER, J. et al. Smartphone ownership and internet usage continues to climb in emerging economies. *Pew Research Center*, Washington, DC, p. 1–44, 2016. Disponível em: <https://www.pewresearch.org>. Acesso em: 18 ago. 2025.

RODRíGUEZ-BARROSOA, N.; LóPEZ, D. J.; LUZóN, M. V.; HERRERA, F.; MARTíNEZ-CáMARA, E. Survey on federated learning threats: concepts, taxonomy on attacks and defences, experimental study and challenges. *arXiv preprint*, 2022. ArXiv:2201.08135. Disponível em: <https://arxiv.org/abs/2201.08135>. Acesso em: 18 ago. 2025.

RUDER, S. *An Overview of Gradient Descent Optimization Algorithms*. 2016. Preprint. ArXiv:1609.04747 [cs.LG]. Disponível em: <https://arxiv.org/abs/1609.04747>. Acesso em: 18 ago. 2025.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural Networks*, v. 61, p. 85–117, 2015.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

SU, L.; XU, J.; YANG, P. A non-parametric view of fedavg and fedprox: Beyond stationary points. *Journal of Machine Learning Research*, v. 24, n. 1, p. 1–48, 2023.

SUMMERS, R. M. Deep learning and computer-aided diagnosis for medical image processing: A personal perspective. In: LU, L.; ZHENG, Y.; CARNEIRO, G.; YANG, L. (Ed.). *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*. Cham, Switzerland: Springer, 2017. p. 3–10. ISBN 978-3-319-42999-1. Disponível em: <https://www.springer.com>. Acesso em: 18 ago. 2025.

TERVEN, J.; CORDOVA-ESPARZA, D.-M.; ROMERO-GONZáLEZ, J.-A.; RAMíREZ-PEDRAZA, A.; CHáVEZ-URBIOLA, E. A. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, v. 58, p. 195, Apr 2025. Disponível em: <https://doi.org/10.1007/s10462-025-11198-7>. Acesso em: 18 ago. 2025.

WANG, H.; WANG, J.; ZHAO, Z.; TAN, Y.; WU, Y.; LIU, H.; YANG, J.; ZHANG, E.; CHEN, X.; RONG, Z.; GUO, S.; LI, Y. Understanding knowledge transferability for transfer learning: A survey. *ACM Computing Surveys*, 2025.

XIAO, H.; RASUL, K.; VOLLGRAF, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

YANG, J.; SHI, R.; WEI, D. et al. Medmnist: A lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, v. 8, n. 1, p. 1–12, 2021.

YANG, Q.; LIU, Y.; CHENG, Y.; KANG, Y.; CHEN, T.; YU, H. *Federated Learning: Challenges, Methods, and Future Directions*. [S.l.]: Morgan & Claypool Publishers, 2019. (Synthesis Lectures on Artificial Intelligence and Machine Learning). Disponível em: <https://doi.org/10.2200/S00974ED1V01Y201912AIM043>. Acesso em: 18 ago. 2025. ISBN 9781681736984.

YASHWANTH, B.; WANG, Z.; DING, D.; WU, J.; HUANG, Q. Adaptive self-distillation for minimizing client drift in heterogeneous federated learning. *Transactions on Machine Learning Research*, 2024.

ZHAO, Y.; LI, M.; LAI, L.; SUDA, N.; CIVIN, D.; CHANDRA, V. Federated learning with non-iid data. *arXiv preprint*, 2018. ArXiv:1806.00582. Disponível em: <https://arxiv.org/abs/1806.00582>. Acesso em: 18 ago. 2025.

ZHU, H.; XU, J.; LIU, S.; JIN, Y. *Federated Learning on Non-IID Data: A Survey*. 2021. Disponível em: <https://arxiv.org/abs/2106.06843>. Acesso em: 18 ago. 2025.