Aprendizado Federado Incremental e Sensível ao Risco para Modelos de Ranqueamento em Cenários com Distribuições Heterogêneas de Dados

Gestefane Rabbi[1], **Celso França (colaborador)**[1],
Thierson Couto Rosa (colaborador)[3], **Jussara M. Almeida (colaborador)**[1],
Daniel Xavier de Sousa (coorientador)[2], **Marcos André Gonçalves (orientador)**[1]

[1]Dep. de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG – Brazil

[2]Instituto Federal de Goiás (IFG) – Goiânia, GO – Brazil

[3]Instituto de Informática (UFG) – Goiânia, GO – Brazil

{gestefane,celsofranca,jussara,mgoncalv}@dcc.ufmg.br
daniel.sousa@ifg.edu.br, thierson@ufg.br

Abstract. *This dissertation introduces FedRisk, a novel Federated Learning to Rank (FLTR) approach designed to address one of the most critical challenges in intelligent distributed information systems: model aggregation under heterogeneous and non-IID data distributions. By integrating a risk-aware aggregation mechanism—capable of weighting client updates according to prediction error variability—with a historical global parameter stabilization strategy, FedRisk promotes both robustness and convergence stability in federated ranking environments. Extensive experiments on the MSLR-WEB10K benchmark demonstrate that FedRisk significantly outperforms strong federated baselines such as FedProx, achieving a 15.6% improvement in nDCG@5 and matching centralized performance in nDCG@10, while substantially reducing performance variance across communication rounds. Beyond empirical gains, the work provides a principled formulation of risk-sensitive aggregation for FLTR, a systematic analysis of heterogeneity effects in federated ranking, and a comprehensive experimental validation under realistic distributed scenarios. By bridging Information Retrieval, Distributed Machine Learning, and privacy-aware system design, this dissertation advances the development of scalable, intelligent, and regulation-aligned Information Systems. Its scientific impact was recognized with the Honorable Mention Award for Best Full Paper at SBBD 2025, underscoring the originality and methodological rigor of the proposed approach.*

Resumo. *Esta dissertação apresenta o FedRisk, uma abordagem inédita de Federated Learning to Rank (FLTR) projetada para enfrentar um dos desafios mais críticos dos Sistemas de Informação distribuídos: a agregação de modelos sob heterogeneidade e distribuições não independentes e não identicamente distribuídas (não-IID). A proposta integra um mecanismo de agregação sensível ao risco — capaz de ponderar as atualizações dos clientes de acordo com a variabilidade do erro de predição — a uma estratégia de estabilização baseada na incorporação de parâmetros globais históricos, promovendo maior robustez e estabilidade no processo de convergência federada.*

Experimentos extensivos no benchmark MSLR-WEB10K demonstram que o FedRisk supera abordagens federadas consolidadas, como o FedProx, alcançando ganho de 15,6% em nDCG@5 e desempenho equivalente ao treinamento centralizado em nDCG@10, além de reduzir substancialmente a variância ao longo das rodadas de comunicação. Para além dos ganhos empíricos, o trabalho apresenta uma formulação principiada de agregação sensível ao risco para FLTR, uma análise sistemática dos efeitos da heterogeneidade no ranqueamento federado e uma validação experimental abrangente em cenários distribuídos realistas. Ao integrar Recuperação de Informação, Aprendizado de Máquina Distribuído e princípios de privacidade e governança de dados, esta dissertação contribui para o avanço de Sistemas de Informação escaláveis, inteligentes e alinhados a requisitos regulatórios contemporâneos. Sua relevância científica foi reconhecida com o prêmio de Honra ao Mérito de Melhor Artigo Completo no SBBD 2025.

1. Introdução

O crescimento exponencial do volume de informações digitais tem redefinido o papel dos Sistemas de Informação na sociedade contemporânea. Aplicações como mecanismos de busca, plataformas de comércio eletrônico, sistemas financeiros e soluções em saúde digital operam hoje em ecossistemas distribuídos, heterogêneos e regulados, nos quais dados são gerados e mantidos em múltiplas organizações e dispositivos. Esse cenário, impulsionado pela computação em nuvem, pela expansão da internet e pela ubiquidade de dispositivos conectados [Mukut et al. 2012], exige arquiteturas mais inteligentes, escaláveis e alinhadas a princípios de privacidade, governança e responsabilidade no uso de dados.

Restrições legais e regulatórias, como legislações de proteção de dados, limitam a centralização de informações para treinamento de modelos de Inteligência Artificial, impondo novos desafios técnicos e organizacionais [Liu et al. 2021]. Nesse contexto, o *Federated Learning* (FL) surge como paradigma promissor para Sistemas de Informação Inteligentes, ao permitir o treinamento colaborativo sobre dados distribuídos sem a necessidade de compartilhamento direto das bases locais, promovendo conformidade regulatória e mitigação de riscos associados à exposição de dados sensíveis [Liu et al. 2021]. Essa característica torna o FL especialmente relevante em domínios críticos, como saúde, finanças e serviços digitais de larga escala [Neto et al. 2020].

Na Recuperação de Informação, a tarefa de *Learning to Rank* (LTR) [Liu 2010] ocupa posição central em sistemas de busca, recomendação e apoio à decisão, pois determina a ordenação de documentos ou itens segundo sua relevância. Entretanto, em ambientes distribuídos, modelos locais de LTR refletem particularidades regionais, comportamentais e contextuais, resultando em distribuições heterogêneas que desafiam abordagens tradicionais de agregação [Yu et al. 2022]. Essa heterogeneidade, típica de Sistemas de Informação reais, pode comprometer tanto a estabilidade quanto a efetividade de modelos globais.

Os métodos de *Federated Learning to Rank* (FLTR) [Li and Ouyang 2021] estendem o LTR ao ambiente federado, mas enfrentam um desafio central: a agregação eficiente de modelos locais sob distribuições não independentes e não identicamente distribuídas (não-IID) [Wang and Zuccon 2022]. Diferenças nos perfis de

usuários, padrões de interação e domínios de consulta geram assimetrias que impactam diretamente a convergência do treinamento. Além disso, a necessidade de reduzir custos de comunicação é determinante para a viabilidade prática dessas soluções [McMahan et al. 2023, Kairouz et al. 2021].

Este trabalho se alinha aos Grandes Desafios de Pesquisa em Sistemas de Informação no Brasil (2016–2026), ao propor agregação federada robusta para ranqueamento sob heterogeneidade e não-IID, com estabilidade e efetividade em cenários distribuídos e sensíveis a dados [Araujo et al. 2017].

Esta dissertação aborda a agregação em FLTR sob heterogeneidade não-IID, propondo mecanismos sensíveis ao risco que equilibram contribuições locais, reduzem instabilidades e promovem ganhos consistentes. A abordagem integra Recuperação de Informação, aprendizado distribuído e governança de dados, contribuindo para Sistemas de Informação que conciliam desempenho, escalabilidade e responsabilidade ética.

Contribuições principais: Esta dissertação apresenta as seguintes contribuições originais: (i) a formulação de um mecanismo de agregação sensível ao risco para FLTR, projetado para cenários não-IID; (ii) a análise sistemática do impacto da heterogeneidade na estabilidade e na efetividade do ranqueamento federado; (iii) a proposição de estratégias para reduzir instabilidades de treinamento mantendo eficiência de comunicação; (iv) uma avaliação experimental abrangente demonstrando ganhos consistentes sobre métodos de agregação tradicionais; (v) a consolidação de evidências que reforçam a viabilidade de arquiteturas federadas como solução estratégica para Sistemas de Informação regulados e orientados por IA.

Parte dos resultados desta dissertação foi consolidada em um artigo completo publicado no Simpósio Brasileiro de Bancos de Dados (SBBD) em 2025, o qual recebeu o prêmio de Honra ao Mérito como melhor artigo completo do evento. Esse reconhecimento evidencia a originalidade metodológica, o rigor experimental e o impacto da pesquisa na comunidade brasileira de Sistemas de Informação e Bancos de Dados.

Contribuição para Sistemas de Informação: Este trabalho contribui ao tratar o processamento distribuído sob restrições de privacidade e governança, viabilizando modelos colaborativos via aprendizado federado sem compartilhamento de dados, em cenários com alta heterogeneidade na distribuição de dados entre os participantes dos treinamentos. Sob a perspectiva de Sistemas de Informação, a proposta viabiliza a implantação de modelos de ranqueamento em ambientes regulados e descentralizados, apoiando processos de tomada de decisão mesmo quando os dados não podem ser centralizados.

2. Definição do Problema

Definições Preliminares Consideramos um cenário de *Federated Learning to Rank* (FLTR) com clientes $\mathcal{K} = \{1, \ldots, K\}$, onde cada cliente k possui dados locais $\mathcal{D}_k = \{(q_i, d_i, y_i)\}_{i=1}^{n_k}$, com rótulos $y_i \in \{0, \ldots, L\}$ e $n_k = |\mathcal{D}_k|$. Cada par (q_i, d_i) é representado por $x_i = \phi(q_i, d_i) \in \mathbb{R}^m$ ($m = 136$ no MSLR-WEB10K). Denotamos por θ_k^t e θ_G^t os modelos local e global na rodada t. Em cada rodada, o servidor seleciona $S_t \subseteq \mathcal{K}$, com $|S_t| = C$, e cada cliente $k \in S_t$ minimiza $F_k(\theta) = \frac{1}{n_k} \sum \ell(\theta; x_i, y_i)$, onde $\ell(\theta; x, y)$ corresponde à função de perda de entropia cruzada (*cross-entropy*) aplicada sobre os rótulos de relevância. O objetivo global é $F(\theta) = \sum_{k \in S_t} \frac{n_k}{n} F_k(\theta)$, com agregação via FedAvg

$\theta_G^t = \sum_{k \in S_t} \frac{n_k}{n} \theta_k^t$. Consideramos cenários não-IID, com desbalanceamento de volume ($n_i \neq n_j$) e distribuição ($P_i(x, y) \neq P_j(x, y)$), induzindo divergência entre modelos locais.

Adicionalmente, definimos um fator de risco por cliente $risk_k^t$, derivado dos erros locais $\mathbf{e}_k^{(b)} = [(\hat{y}_i - y_i)^2]_{i \in b}$, a partir dos quais se obtém $ZRisk(k)$. Esse fator reflete a confiabilidade das atualizações, sendo maior para clientes com erros elevados ou instáveis, cuja influência deve ser reduzida na agregação. A formulação completa é apresentada na Seção 5. Denotamos por $\tilde{\theta}^t$ o modelo agregado após ponderação por risco e por θ_G^{t-1} o modelo global anterior, utilizado como memória histórica.

Formulação do Problema: Em aprendizado federado para ranqueamento, clientes treinam localmente a partir de θ_G^{t-1} e retornam θ_k^t, agregados via FedAvg. Em cenários não-IID, atualizações conflitantes induzem instabilidade, maior variabilidade e ausência de convergência [Wang 2024]. Assim, buscamos melhorar a estabilidade do modelo global sem comprometer o desempenho de ranqueamento.

3. Hipótese e Perguntas de Pesquisa

Esta dissertação propõe o aprimoramento do processo de agregação em *Federated Learning to Rank* (FLTR), considerando que modelos locais treinados em ambientes federados estão sujeitos a diferentes níveis de qualidade, especialmente sob distribuições não-IID. A proposta baseia-se em dois princípios centrais: **sensibilidade ao risco** [Rodrigues et al. 2022] e **memória histórica do modelo global**.

A sensibilidade ao risco, conforme definida em [Rodrigues et al. 2022], busca reduzir a probabilidade de baixo desempenho em consultas individuais mantendo boa eficácia média. Neste trabalho, utilizamos esse conceito para avaliar a confiabilidade dos modelos locais, distinguindo clientes com atualizações mais robustas daqueles mais propensos a ruídos ou desvios. Já a memória histórica refere-se à reutilização controlada dos parâmetros globais das rodadas anteriores, com o objetivo de preservar conhecimento acumulado e mitigar oscilações causadas por atualizações instáveis.

Com base nesses princípios, formulamos duas hipóteses: (i) é possível estimar, a partir dos erros de predição, um indicador de risco que reflita a confiabilidade de cada cliente na rodada corrente; e (ii) a incorporação do histórico do modelo global pode reduzir a variabilidade e aumentar a robustez sob dados não-IID. A primeira hipótese aborda a ausência de mecanismos que identifiquem clientes potencialmente prejudiciais ao modelo global em cenários heterogêneos [Divi et al. 2021], enquanto a segunda enfrenta a limitação recorrente de desconsiderar o conhecimento acumulado nas rodadas anteriores.

Essas hipóteses se desdobram nas seguintes perguntas de pesquisa:

RQ1: *O uso da Sensibilidade ao Risco [Rodrigues et al. 2025], estimada a partir dos erros de predição (MSE [Hastie et al. 2009]), como fator de ponderação na agregação, melhora a eficácia do modelo global?*

RQ2: *A inclusão dos parâmetros do modelo global de rodadas anteriores no processo de agregação reduz a variabilidade e mitiga os efeitos da distribuição não-IID?*

Neste trabalho investigamos se a sensibilidade ao risco na agregação federada melhora o desempenho em cenários não-IID.

4. Fundamentos e Trabalhos Relacionados

Os trabalhos foram selecionados por meio de busca sistemática em ACM Digital Library, IEEE Xplore, SpringerLink e arXiv, utilizando termos como *federated learning*, *learning to rank*, *non-IID data* e *robust aggregation*. Foram priorizados estudos recentes, relevantes e metodologicamente alinhados ao problema, além de trabalhos clássicos para contextualização.

Diversidade nos Modelos de Agregação: Diversas estratégias foram propostas para lidar com cenários não-IID em FL. [Ai et al. 2018] exploram funções de perda baseadas em atenção, enquanto [Zhu et al. 2021] mostram que muitas abordagens priorizam acurácia média, negligenciando estabilidade.

Métodos consolidados incluem FedProx [Li et al. 2020], que introduz regularização proximal; SCAFFOLD [Karimireddy et al. 2020], que utiliza variáveis de controle; e FedNova [Wang and Liu 2020], que normaliza contribuições locais. Outras abordagens, como [Li et al. 2021], alinham representações locais e globais, mantendo agregação baseada em média.

Diferentemente desses métodos, nossa proposta incorpora (i) reutilização do conhecimento histórico do modelo global e (ii) ponderação das contribuições locais com base na sensibilidade ao risco, utilizando qualidade preditiva como critério de agregação no contexto de *Federated Learning to Rank*.

Impacto da Natureza não-IID nos Modelos Federados: A heterogeneidade entre clientes é um dos principais desafios do FL, degradando desempenho e dificultando a convergência [Wang and Zuccon 2022, Jiménez-Gutiérrez et al. 2024, Jiménez-Gutiérrez et al. 2025]. Modelos locais enviesados podem distorcer a agregação e aumentar a variabilidade ao longo das rodadas.

Estudos como [Jiménez-Gutiérrez et al. 2024] classificam diferentes tipos de não-IID (estatístico, sistêmico e temporal), enquanto [Jiménez-Gutiérrez et al. 2025] mostram que variações como *label skew*, *feature skew* e *quantity skew* impactam negativamente métodos tradicionais como o FedAvg.

Nesse contexto, a heterogeneidade deve ser tratada como um obstáculo estrutural, afetando acurácia, estabilidade e eficiência. Nossa abordagem busca mitigar esses efeitos ao reduzir a influência de clientes com baixa confiabilidade e promover maior consistência na evolução do modelo global.

Sensibilidade ao Risco: Em Recuperação de Informação, sensibilidade ao risco refere-se à redução da probabilidade de desempenhos muito baixos em consultas específicas, considerando a variabilidade além da média [Rodrigues et al. 2022, Rodrigues et al. 2025].

No contexto federado, clientes apresentam diferentes níveis de estabilidade preditiva. Trabalhos recentes exploram noções de risco para seleção e ponderação de clientes [Zhao et al. 2024, Ads et al. 2024, Chen et al. 2021], mas não utilizam métricas explícitas de sensibilidade ao risco da literatura de RI nem incorporam diretamente a variabilidade como critério de agregação.

Nossa proposta baseia-se em [Rodrigues et al. 2022, Rodrigues et al. 2025], adaptando a *RiskLoss* para FLTR e utilizando erros de predição para estimar a confia-

bilidade dos clientes. Esses valores são então usados como pesos na agregação global, privilegiando contribuições mais estáveis.

Aproveitamento do Conhecimento Histórico: O uso de conhecimento histórico em FL consiste na reutilização de modelos globais anteriores para estabilizar o treinamento. Essa estratégia pode ser aplicada via regularização local [Lv et al. 2023] ou integração direta na agregação [Wang et al. 2024].

O FedNTD [Lee et al. 2022] utiliza distilação para alinhar previsões entre rodadas, preservando conhecimento de forma indireta. De forma mais próxima, [Hu 2024] exploram o uso de histórico global no treinamento federado.

Diferentemente desses trabalhos, nossa abordagem reutiliza explicitamente o modelo global anterior na agregação, combinando-o com ponderação sensível ao risco. Essa integração fortalece a estabilidade e a robustez em cenários não-IID, otimizando métricas de ranqueamento.

5. Proposta

Para enfrentar a alta variabilidade e melhorar a estabilidade do modelo global em cenários com dados não-IID, propomos duas estratégias complementares. A primeira consiste na introdução de um fator de Sensibilidade ao Risco para ajustar o peso de cada cliente na agregação, considerando a confiabilidade estatística de seu modelo local. A segunda estratégia incorpora o conhecimento histórico ao processo de atualização global, reutilizando parâmetros de rodadas anteriores como mecanismo de estabilização. A seguir, detalhamos cada uma dessas estratégias e sua implementação no contexto de FLTR.

Uso de Sensibilidade ao Risco na Agregação: Na primeira estratégia, substituímos a técnica tradicional de agregação do FL — baseada na proporção do número de amostras locais de cada cliente na rodada ($\frac{n_k}{n}$, conforme o FedAvg) — por uma estratégia que utiliza o fator de Sensibilidade ao Risco [Rodrigues et al. 2022] como ponderador na combinação dos modelos locais.

A motivação para essa substituição é que o volume de dados não necessariamente reflete a qualidade ou a confiabilidade preditiva do modelo local. Em cenários não-IID, clientes com grande número de amostras podem apresentar distribuições restritas de rótulos ou baixa diversidade, comprometendo sua capacidade de generalização. Assim, ponderar exclusivamente por n_k pode amplificar contribuições pouco representativas.

Propomos, portanto, utilizar o complemento do fator estimado para cada cliente, $(1 - risk_k)$, como medida de confiabilidade relativa. A agregação passa a privilegiar clientes com menor propensão a desempenhos significativamente abaixo do esperado.

Formalmente, a agregação ponderada é dada por $\tilde{\theta}^t = \frac{1}{|S_t|} \sum_{k \in S_t} ((1 - risk_k) \cdot \theta_k^t)$ onde S_t é o conjunto de clientes selecionados na rodada t, θ_k^t representa os parâmetros locais do cliente k, e $risk_k$ é o fator de risco estimado com base na variabilidade de seus erros de predição.

A implementação dessa estratégia é ilustrada nos pseudo-códigos a seguir, que descrevem as rotinas executadas no servidor e nos clientes.

No Algoritmo 1, o fluxo padrão do FedAvg é estendido em dois pontos. Após

Algorithm 1: Treinamento Federado Sensível ao Risco (Servidor)

Input: Clientes K, rodadas T, quantidade C
Output: Modelo global θ_G

Inicializar θ_G^0 // Modelo global inicial
for *rodada* $t = 1$ **to** T **do**
  Selecionar subconjunto S_t
  for *cada cliente* $k \in S_t$ **do in parallel**
    $\theta_k^t, risk_k \leftarrow$ **ClientTrain**(k, θ_G^{t-1})
  Agregação Global:
  $\tilde{\theta}^t \leftarrow \frac{1}{|S_t|} \sum_{k \in S_t} (1 - risk_k) \cdot \theta_k^t$
  $\theta_G^t \leftarrow \alpha \cdot \tilde{\theta}^t + \beta \cdot \theta_G^{t-1}$
return θ_G^T
servidor_calcula_riscos(mse_vector):
  return $risks$

Destaques em vermelho indicam as alterações propostas.

Algorithm 2: ClientTrain (Cliente)

Output: $\theta_k, risk_k$
$\theta_k \leftarrow \theta$
$\mathcal{R}_k \leftarrow \emptyset$
for *época* $e = 1$ **to** E **do**
  for *batch* $b = 1$ **to** B **do**
    Obter Y_k^b
    $outputs \leftarrow net(X_k^b)$
    $P_k^b \leftarrow \arg\max(outputs)$
    $mse\_vector_k^b = [(p_i - y_i)^2 \mid \forall i \in b]$
    $risks \leftarrow$ servi-dor_calcula_risks$(mse_vector_k^b)$
    $risk_k^b \leftarrow risks[k]$
    $\mathcal{R}_k \leftarrow \mathcal{R}_k \cup \{risk_k^b\}$
    Otimizar modelo local θ_k
if $|\mathcal{R}_k| > 0$ **then**
  $risk_k \leftarrow$ mediana$(\mathcal{R}_k)$
else
  $risk_k \leftarrow 0$
return $\theta_k, risk_k$

Notação: X_k^b: features, P_k^b: predições, Y_k^b: rótulos, $risk_k^b$: risco.

a seleção do subconjunto S_t na rodada t, o servidor executa em paralelo o Algoritmo 2. Diferentemente do FedAvg tradicional, cada cliente retorna não apenas os parâmetros locais θ_k^t, mas também um escalar $risk_k$. Esses valores são utilizados para ponderar as contribuições locais conforme a Equação (5), substituindo diretamente a média ponderada por $\frac{n_k}{n}$. Em seguida, o modelo agregado corrente $\tilde{\theta}^t$ é combinado com o modelo global da rodada anterior, incorporando memória histórica ao processo de atualização.

No Algoritmo 2, a rotina **ClientTrain** é estendida para calcular e retornar o risco associado a cada cliente. Após inicializar o modelo local $\theta_k \leftarrow \theta$, o cliente mantém uma estrutura para armazenar riscos por *batch*. Durante o treinamento, para cada *batch*, o cliente calcula o vetor de erros quadráticos médios $mse\_vector_k^b = \left[(p_i - y_i)^2 \mid i \in b\right]$, comparando as predições p_i com os rótulos verdadeiros y_i. Esse vetor é enviado ao servidor para cálculo do fator de sensibilidade ao risco, conforme detalhado na Seção 5.1.1. Ao final das épocas locais, o risco final $risk_k$ é obtido pela mediana dos riscos acumulados e retornado ao servidor juntamente com θ_k^t.

Essa integração entre cliente e servidor permite incorporar sensibilidade ao risco diretamente na etapa de agregação global, ajustando dinamicamente o peso das contribuições locais de acordo com sua confiabilidade estatística.

Definição de risco no contexto de ranqueamento federado: Em recuperação de informação, o risco é tipicamente definido no nível de consulta, refletindo a variabilidade entre queries [Rodrigues et al. 2022, Rodrigues et al. 2025]. No entanto, em cenários de aprendizado federado, a agregação ocorre no nível de cliente. Dessa forma, definimos o risco no nível do cliente, baseado nos erros locais, de modo a capturar a confiabilidade das atualizações em cenários não-IID e manter alinhamento com o objetivo de otimização da agregação federada.

Cálculo do Fator de Risco: Inspirados na formulação proposta por [Rodrigues et al. 2022, Rodrigues et al. 2025], adaptamos o cálculo da Sensibilidade ao Risco para o contexto federado, considerando os erros de predição observados

nos clientes participantes da rodada t. Seja $S_t \subseteq \{1, \ldots, K\}$ o conjunto de clientes selecionados na rodada corrente, com $|S_t|$ clientes. Para cada cliente $k \in S_t$, consideramos um *batch* local com n_b instâncias (x_i, y_i), onde $x_i = \phi(q_i, d_i)$ representa o vetor de atributos extraído do par consulta-documento (q_i, d_i) e y_i é o respectivo rótulo de relevância. Cada cliente calcula localmente suas predições $p_{ki} = f_k(x_i)$ e constrói o vetor de erros quadráticos do *batch* como $MSE_k = \left[(p_{k1} - y_{k1})^2, \ldots, (p_{kn_b} - y_{kn_b})^2\right]$. A partir dos vetores MSE_k, construímos a matriz de erros $M \in \mathbb{R}^{|S_t| \times n_b}$, cujas entradas são dadas por $m_{ki} = (p_{ki} - y_{ki})^2$. Definimos então: (i) o somatório total dos erros $N = \sum_{k=1}^{|S_t|} \sum_{i=1}^{n_b} m_{ki}$; (ii) o somatório por cliente $L_k = \sum_{i=1}^{n_b} m_{ki}$; (iii) o somatório por instância $T_i = \sum_{k=1}^{|S_t|} m_{ki}$. O erro esperado para cada célula da matriz é calculado como $e_{ki} = \frac{L_k \cdot T_i}{N}$, e o desvio padronizado do erro observado em relação ao esperado é $z_{ki} = \frac{m_{ki} - e_{ki}}{\sqrt{e_{ki}}}$.

Seguindo [Rodrigues et al. 2022], agregamos esses desvios na métrica $ZRisk(k)$: $ZRisk(k) = \sum_{i \in J^-} z_{ki} + (1 + \alpha) \sum_{i \in J^+} z_{ki}$, onde $J^+ = \{i \mid z_{ki} \geq 0\}$ e $J^- = \{i \mid z_{ki} < 0\}$. O parâmetro α controla o grau de aversão ao risco, penalizando mais fortemente desvios positivos (erros acima do esperado). Como $ZRisk(k)$ não incorpora explicitamente a magnitude média dos erros, utilizamos a métrica GeoRisk [Dincer et al. 2016], que combina risco estatístico e erro médio por meio de uma média geométrica: $GeoRisk(k) = \sqrt{\left(\frac{1}{n_b} \sum_{i=1}^{n_b} m_{ki}\right) \cdot \Phi\left(\frac{ZRisk(k)}{n_b}\right)}$, onde $\Phi(\cdot)$ representa a função de distribuição acumulada da normal padrão.

Por fim, definimos o fator de sensibilidade ao risco do cliente k comparando seu risco ao de um sistema ideal Z, definido como a média dos erros por instância $Risk_k = GeoRisk(Z) - GeoRisk(k)$, em que $Z_i = \frac{1}{|S_t|} \sum_{k=1}^{|S_t|} m_{ki}$ representa o erro médio por coluna da matriz M. O valor $Risk_k$ é então utilizado como fator de ponderação na agregação global, conforme descrito na Seção 5.1, reduzindo a influência de clientes com maior variabilidade e maior propensão a erros extremos.

Adição Incremental do Modelo Global na Agregação: A abordagem proposta busca incorporar explicitamente os parâmetros do modelo global da rodada anterior ao processo de agregação, combinando-os com o modelo obtido na rodada corrente após a ponderação por sensibilidade ao risco.

Seja $\tilde{\theta}^t$ o vetor de parâmetros resultante da agregação ponderada dos modelos locais na rodada t, e θ_G^{t-1} o modelo global da rodada anterior. O novo modelo global é então calculado por meio da seguinte combinação linear:

$$\theta_G^t = \alpha \cdot \tilde{\theta}^t + \beta \cdot \theta_G^{t-1} \tag{1}$$

onde $\alpha, \beta \in \mathbb{R}^+$ são hiperparâmetros que controlam, respectivamente, a influência do modelo agregado atual e do modelo histórico no processo de atualização.

Diferentemente da agregação tradicional baseada exclusivamente nas contribuições locais da rodada corrente, essa estratégia introduz uma componente de memória temporal diretamente na etapa de combinação global. A reutilização de θ_G^{t-1} atua como mecanismo de estabilização, reduzindo oscilações decorrentes da heterogeneidade não-IID e contribuindo para uma dinâmica de atualização mais suave ao longo das rodadas.

6. Experimentos

Configuração Experimental: A configuração experimental foi estruturada de modo a refletir os desafios observados em cenários reais de aprendizado federado, particularmente na tarefa de aprendizado para ranqueamento.

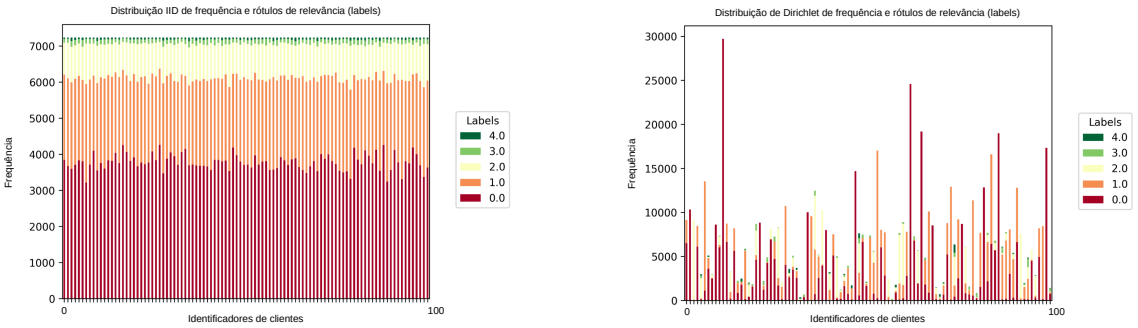
Configuração do treinamento local: O treinamento local utilizou gradiente descendente estocástico com taxa de aprendizado de `3.0e-4`, uma época local por rodada e *batch size* de 16 ou 32. O sistema operou com 100 clientes, fração de participação de 10% por rodada e 100 rodadas no total. Esses parâmetros foram mantidos constantes entre os métodos de agregação avaliado. A menos que especificado de outra forma, adotamos $\alpha = 1$ e $\beta = 1$, por apresentarem o melhor desempenho empírico ao longo dos experimentos.

Dados e Biblioteca: Os experimentos utilizaram o dataset *MSLR-WEB10K* [Qin and Liu 2013], um *benchmark* amplamente adotado em aprendizado para ranqueamento [Köppel et al. 2019], composto por aproximadamente 10.000 consultas associadas a múltiplos documentos, representados por vetores de 136 atributos numéricos extraídos de páginas da web. Para simular o ambiente federado, foi utilizada a biblioteca *Flower* [Beutel et al. 2020], uma estrutura de código aberto voltada à implementação e execução de sistemas de aprendizado federado de forma escalável. No contexto deste trabalho, o *Flower* foi empregado em modo de simulação local, permitindo a criação de múltiplos clientes em um único ambiente físico, com controle sobre a distribuição dos dados, o número de rodadas e o particionamento dos conjuntos de treino e validação.

Utilizamos o experimento com dados centralizados como *baseline*, permitindo a comparação direta com os desempenhos nos ambientes de aprendizado federado. Diferente do cenário federado, onde os dados (pares consulta-documento) são distribuídos entre múltiplos clientes, neste experimento os dados de treino e de avaliação permaneceram centralizados e foram utilizados integralmente para o treinamento e avaliação de um único modelo.

Arquitetura dos modelos de ranqueamento: Os modelos utilizam uma rede multicamadas que prediz rótulos de relevância (5 classes) a partir de 136 características. A arquitetura é dada por $\text{logits} = W_2\,\text{ReLU}(W_1 x + b_1) + b_2$, onde $x \in \mathbb{R}^{136}$, a camada oculta possui 64 unidades e a saída gera 5 logits. Os logits são otimizados diretamente com entropia cruzada durante o treinamento local.

Distribuição dos Dados IID e não-IID: Para simular diferentes cenários no aprendizado federado, usamos a biblioteca *Flower* [Beutel et al. 2020, Yurochkin et al. 2019], que implementa a técnica de particionamento baseada na distribuição *Dirichlet*, com $\alpha = 0.5$, o que induz um cenário de *label skew* controlado entre as partições locais. A Figura 1 ilustra a alocação dos dados entre 100 clientes, dos quais 10 são selecionados aleatoriamente em cada rodada de treinamento. No cenário IID (Figura 1a), observa-se que todos os clientes possuem aproximadamente o mesmo número total de instâncias, e a distribuição das classes (0 a 4) dentro de cada cliente é proporcional à distribuição global do conjunto de dados, ou seja, todas as classes estão igualmente representadas em cada partição (cliente). Em contraste, no cenário não-IID (Figura 1b), verifica-se uma significativa heterogeneidade, com alguns clientes concentrando dados em uma ou poucas classes, enquanto outros apresentam distribuições bastante desbalanceadas.



(a) Dados IID **(b) não-IID**

Figura 1. Distribuições IID e não-IID de dados entre os 100 clientes.

Metodologia de Avaliação: Para mitigar vieses de particionamento, utilizamos validação cruzada com 5 *folds* [Berrar 2018], garantindo que os resultados reflitam o desempenho médio do modelo. No contexto federado, essa estratégia é essencial para avaliar robustez, estabilidade e consistência diante de dados não-IID.

Para avaliar a efetividade dos modelos propostos, especialmente em tarefas de ranqueamento, utilizam-se métricas que consideram não apenas a acurácia dos itens retornados, mas também sua posição nas listas geradas. A métrica nDCG@k (*Normalized Discounted Cumulative Gain*) é amplamente adotada por considerar tanto a relevância quanto a posição dos documentos, atribuindo maiores pesos aos itens relevantes nas primeiras posições — aspecto crucial em cenários onde os usuários interagem principalmente com os primeiros resultados [Järvelin and Kekäläinen 2002, Wang et al. 2013]. Já a métrica MRR (*Mean Reciprocal Rank*) mede a posição do primeiro item relevante, sendo útil para avaliar a capacidade do sistema de retornar boas recomendações no topo do *ranking* [Voorhees 1999].

Essas métricas são indicativas da qualidade do ranqueamento, sendo que valores mais altos refletem melhor desempenho. Para os resultados dos cinco *folds*, foram estimados intervalos de confiança de 95%, permitindo quantificar a variabilidade e conferir rigor estatístico às análises comparativas entre os métodos propostos e os *baselines*.

Reprodutibilidade: O código será disponibilizado publicamente após a aceitação para garantir total reprodutibilidade.

7. Resultados

Avaliação com o método proposto - FedRisk: Avaliamos o método proposto, **FedRisk**, em comparação com métodos de agregação *baselines* da literatura. Para cada estratégia de agregação avaliada, treinamos os modelos separadamente em cada *fold*. Avaliamos, nos dados de teste, o desempenho de cada modelo resultante do treinamento por *fold* e tomamos a média destes desempenhos. Esses resultados estão reportados na Tabela 1. Esta tabela apresenta uma coluna com o número de clientes por rodada de treinamento (10 selecionados aleatoriamente de 100 possibilidades) e, nas colunas seguintes, os valores de desempenho médios, ao final das 100 rodadas, nas métricas nDCG(@1,@5,@10) e MRR(@1,@5,@10), juntamente com seus respectivos intervalos de confiança (valores entre parênteses)[1].

[1]Utilizamos as métricas nDCG e MRR nas posições 1, 5 e 10, que são amplamente reconhecidas na literatura para medir a qualidade de sistemas de ranqueamento [Järvelin and Kekäläinen 2002, Voorhees 1999].

#	Estratégia de Agregação	Clientes por rodada	nDCG (x100)			MRR (x100)		
			@1	@5	@10	@1	@5	@10
1	**FedRisk**	10	27.0 (0.8)	**31.8 (1.0)**	**37.3 (1.2)**	**48.0 (1.5)**	**63.9 (1.3)**	**64.9 (1.3)**
2	FedAvgM	10	**27.4 (11.3)**	29.8 (10.0)	34.2 (8.5)	46.4 (16.2)	61.6 (14.0)	62.8 (13.2)
3	FedProx (μ=0.9)	10	24.1 (6.0)	27.5 (4.1)	32.8 (3.6)	41.8 (8.2)	58.2 (7.5)	59.5 (7.1)
4	FedOpt	10	24.1 (9.5)	27.5 (7.4)	32.7 (7.0)	42.6 (14.3)	58.5 (11.6)	59.9 (11.0)
5	FedTrimmedAvg	10	20.8 (5.3)	26.2 (5.1)	32.5 (4.6)	37.3 (9.4)	54.9 (8.4)	56.4 (7.9)
6	FedAvg	10	23.0 (8.1)	26.2 (7.0)	31.6 (6.7)	39.8 (11.7)	56.0 (11.2)	57.5 (10.5)
7	FedAdam	10	17.8 (3.3)	22.8 (3.5)	28.8 (3.5)	32.9 (5.5)	50.5 (5.9)	52.3 (5.4)
8	FedMedian	10	17.9 (3.0)	22.6 (3.4)	28.3 (3.8)	33.1 (4.9)	50.3 (5.6)	52.0 (5.2)
9	FedYogi	10	17.0 (1.1)	21.3 (1.2)	27.1 (1.2)	31.0 (1.7)	47.8 (2.0)	49.8 (1.9)
10	FedAdagrad	10	16.7 (2.0)	20.9 (1.4)	26.7 (1.6)	30.8 (2.7)	47.3 (2.6)	49.3 (2.5)
11	Centralizado	–	32.4 (3.9)	35.0 (3.2)	35.9 (2.2)	58.5 (5.7)	73.2 (4.1)	73.7 (4.0)

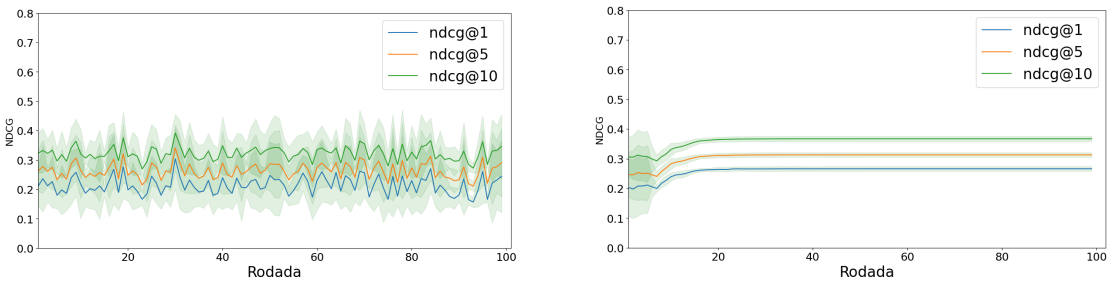
Tabela 1. Comparação do desempenho entre modelos em cenários com dados não-IID (linhas 1–10) na MSLR-WEB10K. Resultados multiplicados por 100, com 5-folds e intervalo de confiança de 95% ao longo de 100 rodadas de treinamento.

Como mostrado na Tabela 1, em cinco das seis métricas avaliadas nos cenários com dados não-IID, o nosso método, **FedRisk**, superou todas as estratégias de agregação consideradas nos experimentos. Em nDCG@5, por exemplo, o **FedRisk** alcançou **31.8**, enquanto o FedProx, que foi o melhor *baseline* entre os comparados (considerando efetividade e estabilidade), atingiu 27.5, um ganho de **15.6%**. Em relação ao FedAdagrad, na mesma métrica, o ganho foi ainda maior, **52.1%**.

Por fim, o **FedRisk** foi o modelo que mais se aproximou do desempenho do modelo centralizado, destacando-se em nDCG@5 e nDCG@10, inclusive superando-o nesta última métrica. Esse resultado é especialmente relevante no cenário não-IID considerado.

Análise da Variabilidade e Convergência: No aprendizado federado, a natureza não-IID dos dados acentua a importância de analisar a variabilidade dos resultados, devido às flutuações entre rodadas e *folds*. Como observa Spiegelhalter [Spiegelhalter 2024], métricas com boas médias podem mascarar alta incerteza e baixa consistência quando há grande variabilidade, comprometendo a confiabilidade. Um exemplo é o FedAvgM na Tabela 1: apesar dos bons valores médios em nDCG@1, o IC elevado revela instabilidade entre particionamentos e baixa consistência.

Utilizamos intervalos de confiança sobre os resultados entre os *folds* em cada rodada para evidenciar a variabilidade dos modelos FedProx e **FedRisk** [Brownlee 2018]. As sombras nas Figuras 2-a e 2-b representam essas amplitudes. Observa-se na Figura 2-b que o **FedRisk** reduz significativamente a variabilidade entre os cinco *folds*, com intervalos mais estreitos em todas as curvas. Já a Figura 2-a evidencia a alta variabilidade do FedProx. Além disso, o **FedRisk** converge rapidamente, atingindo máxima performance e estabilidade em menos de 20 rodadas, enquanto o FedProx não apresenta convergência clara.



(a) - Agregação com FedProx. **(b) - Agregação com FedRisk**

Figura 2. **Aplicação de FedRisk em proporções máximas ($\alpha = 1.0, \beta = 1.0$) do modelo com sensibilidade ao risco e do modelo anterior. — 10 clientes por rodada.**

Os resultados indicam que a combinação das estratégias promove ganhos consistentes em desempenho e estabilidade, posicionando o **FedRisk** como a abordagem mais robusta e respondendo positivamente às RQ1 e RQ2.

Foi utilizado o teste de Wilcoxon pareado unilateral, com tamanho de efeito $r = |Z|/\sqrt{N}$ [Field 2013, Lakens 2013][2]. Em relação ao FedProx, houve significância em nDCG@5 e nDCG@10 ($p = 0.03125$, $r = 0.9045$), enquanto em MRR@5 e MRR@10 não houve significância ($p = 0.09375$), apesar de efeito grande ($r = 0.6633$). Considerando todos os baselines, sob a análise conjunta de significância (Wilcoxon) e magnitude (*effect size*), o **FedRisk** apresentou ganhos consistentes em todas as métricas, com predominância de efeitos grandes, indicando superioridade prática mesmo quando $p \geq 0.05$, possivelmente devido ao baixo número de *folds*.

Análise da Contribuição dos Componentes da Solução: Investigamos como o desempenho do **FedRisk** é influenciado pelos pesos de seus dois componentes: o modelo médio dos clientes na rodada atual ($\tilde{\theta}_t$) e o modelo global da rodada anterior (θ_G^{t-1}), conforme a Equação 1. Para isso, variamos os pesos α e β em diferentes proporções, explorando diferentes cenários de influência dos componentes.

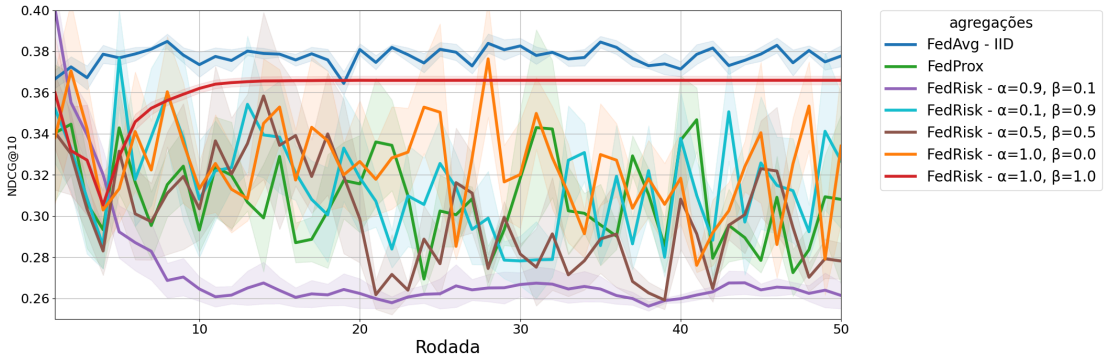


Figura 3. **Comparação entre diferentes proporções de cada componente na composição do modelo global, usando 5 clientes por rodada.**

A Figura 3 apresenta os resultados da análise, comparando as configurações em termos da métrica nDCG@10 ao longo das rodadas, incluindo dois modelos de referência:

[2]O tamanho de efeito r é derivado da estatística Z do teste de Wilcoxon, normalizada pelo tamanho da amostra N, sendo amplamente utilizado como medida de magnitude em testes não paramétricos.

o FedAvg com dados IID e o *baseline* FedProx. Note que o cenário de simplesmente "desligar"o componente $\tilde{\theta}^t$, fazendo $\alpha = 0$, não foi considerado pois é equivalente a desconsiderar o aprendizado e manter o mesmo modelo global por todas as rodadas.

Nossa análise de componentes considerou as seguintes configurações:

- ($\alpha = 0.9$, $\beta = 0.1$): Maior peso do modelo agregado e menor peso do modelo anterior privilegia o aprendizado dos clientes. A curva instável indica que atribuir maior peso a agregação por risco não estabiliza o desempenho do modelo global nem reduz a variabilidade.
- ($\alpha = 0.1$, $\beta = 0.9$): Maior contribuição do modelo anterior e menor do agregado enfraquece o aprendizado com base nos clientes. A curva indica que confiar mais na memória histórica não melhora a efetividade do modelo global nem reduz a variabilidade.
- ($\alpha = 0.5$, $\beta = 0.5$): Mesmo peso para as contribuições do modelo agregado e do modelo anterior. A curva sugere que o equilíbrio entre memória e aprendizado atual sem considerar o máximo de importância para os dois fatores ainda não é suficiente para alcançar os melhores resultados.
- ($\alpha = 1.0$, $\beta = 0.0$): Peso total do modelo baseado no risco e nenhum da memória prioriza apenas o aprendizado dos clientes, mas a curva mostra que isso não melhora o desempenho nem reduz a variabilidade.
- ($\alpha = 1.0$, $\beta = 1.0$): Peso total aos dois componentes, produzindo uma curva altamente estável, *com valores próximos do do FedAvg com dados IID e muito próximo do ideal*. **Esta configuração equilibra eficácia e consistência**, produzindo uma curva estável com desempenho próximo ao FedAvg em cenário IID. Ao combinar aprendizado local e memória global, evita desvios causados por clientes de baixa qualidade e valoriza atualizações consistentes, resultando em maior estabilidade e robustez.

Esta análise mostra que as duas estratégias propostas são igualmente importantes, complementares e necessárias. O melhor cenário é obtido quando ambas têm peso igual a um (valor máximo).

Limitações: Os experimentos foram conduzidos em ambiente controlado, não capturando aspectos reais como latência, falhas e dinâmica dos dados. Assim, os resultados indicam validade interna, sendo necessária validação em cenários reais.

8. Conclusões

Esta dissertação aborda o desafio da agregação em *Federated Learning to Rank* sob não-IID, propondo o **FedRisk** como uma abordagem sensível ao risco para maior estabilidade. A combinação de ponderação baseada em erro local e reutilização da memória global reduz instabilidades e fortalece o treinamento federado.

Os resultados experimentais evidenciam ganhos consistentes de efetividade e estabilidade: o desempenho em *nDCG@5* aumentou de **27.5** (FedProx) para **31.8** (**FedRisk**), representando ganho relativo de **15.6%**, enquanto a variabilidade em *nDCG@1* foi reduzida de **11.3** para **0.8** quando comparado ao FedAvgM. Além disso, o FedRisk igualou o desempenho do modelo centralizado em *nDCG@10*, resultado inédito entre os *baselines* avaliados no cenário não-IID. A análise isolada e combinada dos componentes confirmou que as estratégias propostas são complementares e fundamentais para aumentar a efetividade, reduzir a variabilidade e acelerar a convergência.

Sob a perspectiva de Sistemas de Informação, os resultados indicam que arquiteturas federadas sensíveis ao risco são uma alternativa robusta para ambientes distribuídos e regulados, conciliando desempenho, escalabilidade e uso responsável dos dados.

Embora os experimentos tenham sido conduzidos em ambiente simulado, o *setup* adotado reflete características centrais de sistemas federados reais, incluindo participação parcial de clientes, comunicação em rodadas e distribuições não-IID.

Trabalhos Futuros: As perspectivas incluem: adaptação do FedRisk para cenários com participação dinâmica de clientes; integração com *curriculum learning* e seleção de instâncias para acelerar convergência e reduzir custo; avaliação com métodos recentes de ranqueamento federado *online*, como o FedOLTR; e extensão para domínios como recomendação, recuperação multimodal, saúde digital e finanças. Essas direções ampliam o alcance interdisciplinar da proposta, conectando Recuperação de Informação, aprendizado distribuído e governança de dados, e consolidam o FedRisk como base para Sistemas de Informação federados mais robustos e escaláveis.

Referências

Ads, Z. et al. (2024). Risk-aware accelerated federated learning over heterogeneous wireless networks. *arXiv preprint arXiv:2401.09267*.

Ai, Q., Bi, K., Guo, J., and Croft, W. B. (2018). Learning a deep listwise context model for ranking refinement. In *ACM SIGIR Conference*. ACM.

Araujo, R. M. d., Maciel, R. S. P., and Boscarioli, C. (2017). I GranDSI-BR: Grandes Desafios de Pesquisa em Sistemas de Informação no Brasil (2016–2026). Technical report, Comissão Especial de Sistemas de Informação (CE-SI), Sociedade Brasileira de Computação (SBC). Relatório Técnico.

Berrar, D. (2018). *Cross-Validation*.

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Kwing, H. L., Parcollet, T., Gusmão, P. P. d., and Lane, N. D. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.

Brownlee, J. (2018). *Statistical Methods for Machine Learning*. Machine Learning Mastery.

Chen, S. et al. (2021). Risk-aware federated learning in crowdsensing systems. *arXiv preprint arXiv:2101.01266*.

Dincer, B., Zhu, Y., Craswell, N., and Zhang, M. (2016). Risk-sensitive evaluation and learning to rank using multiple baselines. In *ACM SIGIR*, pages 483–492.

Divi, S., Lin, Y.-S., Farrukh, H., and Celik, Z. B. (2021). New metrics to evaluate the performance and fairness of personalized federated learning.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications, 4 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hu, C. (2024). Improving federated learning accuracy with the incremental averaging method: A comparative analysis of model aggregation techniques. In *Applied and Computational Engineering*, pages 150–157. EWA Publishing.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Jiménez-Gutiérrez, D. M., Hassanzadeh, M., Anagnostopoulos, A., et al. (2025). A thorough assessment of the non-iid data impact in federated learning. Available at: https://arxiv.org/abs/2503.17070.

Jiménez-Gutiérrez, D. M., Solans, D., Heikkilä, M., et al. (2024). Non-iid data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. Available at: https://arxiv.org/abs/2411.12377.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in ML*, 14(1–2):1–210.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S. U., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *ICML*.

Köppel, M., Segner, A., Wagener, M., Pensel, L., Karwath, A., and Kramer, S. (2019). Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. *arXiv preprint arXiv:1909.02768*.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, Volume 4 - 2013.

Lee, G., Jeong, M., Shin, Y., Bae, S., and Yun, S.-Y. (2022). Preservation of the global knowledge by not-true distillation in federated learning.

Li, C. and Ouyang, H. (2021). Federated unbiased learning to rank.

Li, Q., He, B., and Song, D. (2021). Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, pages 429–450.

Liu, J., Huang, J., et al. (2021). From distributed machine learning to federated learning: A survey. Available at: `https://arxiv.org/abs/2104.14362`.

Liu, T.-Y. (2010). Learning to rank for information retrieval. In *ACM SIGIR*, page 904.

Lv, Y., Ding, H., Wu, H., Zhao, Y., and Zhang, L. (2023). Fedrds: Federated learning on non-iid data via regularization and data sharing. *Applied Sciences*, 13(23).

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

Mukut, S., Kakoli, G., and Jyotika, B. (2012). Federated search: An information retrieval strategy for scholarly literature.

Neto, H. N. C., Mattos, D. M. F., and Fernandes, N. C. (2020). Privacidade do usuário em aprendizado colaborativo: Federated learning, da teoria à prática. In *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSEG)*.

Qin, T. and Liu, T. (2013). Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597.

Rodrigues, P. H., Danielde Sousa, França, C., Rabbi, G., Rosa, T., and Gonçalves, M. A. (2025). Risk-sensitive optimization of neural deep learning ranking models with applications in ad-hoc retrieval and recommender systems. *IP&M*, 62(4):104126.

Rodrigues, P. H. S., Xavier Sousa, D., Couto Rosa, T., and Gonçalves, M. A. (2022). Risk-sensitive deep neural learning to rank. In *ACM SIGIR*, page 803–813.

Spiegelhalter, D. (2024). *The Art of Uncertainty: How to Navigate Chance, Ignorance, Risk and Luck*. Pelican Books.

Voorhees, E. M. (1999). The trec-8 question answering track report. In *TREC-8*. National Institute of Standards and Technology (NIST).

Wang, H., Xu, H., Li, Y., Xu, Y., Li, R., and Zhang, T. (2024). FedCDA: Federated learning with cross-rounds divergence-aware aggregation. In *ICLR*.

Wang, J. and Liu, M. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. In *NeurIPS*.

Wang, S. (2024). Effective and secure federated online learning to rank. *arXiv preprint arXiv:2412.19069*.

Wang, S. and Zuccon, G. (2022). Is non-iid data a threat in federated online learning to rank? In *ACM SIGIR Conference*, SIGIR '22, page 2801–2813.

Wang, Y., Li, T.-Y., Wang, D., and Zhu, M. (2013). A theoretical analysis of ndcg type ranking measures. *Journal of Machine Learning Research*, 14:25–54.

Yu, T., Bagdasaryan, E., and Shmatikov, V. (2022). Salvaging federated learning by local adaptation.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, T. N., and Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks.

Zhao, S. et al. (2024). Federated risk-aware learning with central sensitivity estimation. *arXiv preprint arXiv:2502.17694*.

Zhu, H., Jin, B., Li, H., and Liang, X. (2021). Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390.