Aprendizado Federado Incremental e Sensível ao Risco para Modelos de Ranqueamento

em Cenários com Distribuições Heterogêneas de Dados

Gestefane Rabbi Magalhães[1], Daniel Xavier de Sousa[2] (Coorientador), Marcos André Gonçalves[1] (Orientador)

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Brasil

[2]Instituto Federal de Goiás (IFG) – Brasil

{gestefane, mgoncalv}@dcc.ufmg.br, daniel.sousa@ifg.edu.br

Resumo. *Esta dissertação propõe o FedRisk, um novo método de Federated Learning to Rank (FLTR) desenvolvido para lidar com um dos principais obstáculos do aprendizado federado: a construção de um modelo global confiável quando os dados dos participantes apresentam elevada heterogeneidade e seguem distribuições não independentes e não identicamente distribuídas (não-IID). O método combina uma estratégia de agregação orientada ao risco, que ajusta a contribuição de cada cliente com base na variação de seus erros de predição, com um mecanismo de estabilização que reutiliza parâmetros globais produzidos em rodadas anteriores. Essa integração busca reduzir oscilações durante o treinamento e tornar a convergência federada mais estável e robusta. Uma ampla avaliação experimental no benchmark MSLR-WEB10K evidencia que o FedRisk supera métodos federados amplamente adotados, entre eles o FedProx, obtendo melhoria de 15,6% em nDCG@5 e resultados comparáveis aos do treinamento centralizado em nDCG@10, além de diminuir de forma expressiva a variabilidade observada ao longo das rodadas de comunicação. Para além do desempenho alcançado, a pesquisa introduz uma formulação fundamentada de agregação sensível ao risco aplicada a FLTR, examina de maneira sistemática os impactos da heterogeneidade de dados sobre o ranqueamento federado e realiza uma validação experimental abrangente em configurações distribuídas representativas. Abordando temas em Recuperação de Informação, Aprendizado de Máquina Distribuído e requisitos relacionados à privacidade de dados, essse trabalho amplia o conhecimento sobre o desenvolvimento de modelos de ranqueamento federados mais eficazes, estáveis e adequados a ambientes descentralizados. A relevância científica da pesquisa foi reconhecida pela concessão da Honra ao Mérito de Melhor Artigo Completo no SBBD 2025.*

Abstract. *This dissertation proposes FedRisk, a novel Federated Learning to Rank (FLTR) method designed to address one of the major challenges of federated learning: constructing a reliable global model when client data are highly heterogeneous and follow non-independent and non-identically distributed (non-IID) distributions. The proposed method combines a risk-aware aggregation strategy, which adjusts each client's contribution according to the variability of its prediction errors, with a stabilization mechanism that reuses historical global parameters from previous communication rounds. This*

combination reduces training oscillations and promotes a more stable and robust federated convergence process. An extensive experimental evaluation on the MSLR-WEB10K benchmark shows that FedRisk outperforms widely adopted federated methods, including FedProx, achieving a 15.6% improvement in nDCG@5 while reaching performance comparable to centralized training in terms of nDCG@10, in addition to substantially reducing performance variability throughout the communication rounds. Beyond these empirical gains, the proposed approach introduces a principled risk-aware aggregation framework for FLTR, systematically investigates the effects of data heterogeneity on federated ranking, and provides a comprehensive experimental validation under representative distributed settings. By addressing challenges in Information Retrieval, Distributed Machine Learning, and data privacy, this work advances the development of more effective, stable, and reliable federated ranking models for decentralized environments. The scientific relevance of this research was recognized with the Honorable Mention Award for Best Full Paper at SBBD 2025.

Categoria: Mestrado
Programa: Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Minas Gerais (UFMG)
Defendida e aprovada em: 14 de outubro de 2025

Banca examinadora:

- Marcos André Gonçalves (Departamento de Ciência da Computação (DCC) - Universidade Federal de Minas Gerais (UFMG))
- Daniel Xavier de Sousa (Instituto Federal de Goiás (IFG))
- Idilio Drago (Computer Science Department (CSD) - University of Turin)
- Heitor Soares Ramos Filho (Departamento de Ciência da Computação (DCC) - Universidade Federal de Minas Gerais (UFMG))

Nota: Aprovado
Escala de avaliação: 0–100

Destaques:

- Agregação sensível ao risco melhora a estabilidade do FLTR sob dados não-IID
- O FedRisk aproxima o desempenho de modelos centralizados em cenários de ranqueamento
- Uso de memória global reduz a variabilidade entre rodadas federadas
- Ganho de 15,6% em nDCG@5 em relação a um forte baseline federado
- Suporte a ranqueamento com privacidade em sistemas distribuídos

Contexto e problema: O crescimento exponencial do volume de informações digitais não apenas transformou, mas reconfigurou estruturalmente os Sistemas de Informação, que passaram a operar em ambientes altamente distribuídos, heterogêneos e sujeitos a restrições regulatórias cada vez mais rigorosas.

Aplicações críticas — como mecanismos de busca, sistemas de recomendação, plataformas de comércio eletrônico e infraestruturas digitais em larga escala — dependem diretamente da capacidade de aprender a partir de dados distribuídos, frequentemente impossíveis de serem centralizados devido a restrições técnicas, legais e operacionais.

Nesse cenário, o aprendizado federado emerge como uma abordagem fundamental para viabilizar aprendizado em ambientes regulados, permitindo treinamento colaborativo sem compartilhamento direto dos dados e conciliando privacidade, governança e eficiência [Neto et al. 2020]. Entretanto, sua aplicação prática é limitada por um dos principais gargalos atuais do aprendizado federado: a presença de distribuições não independentes e não identicamente distribuídas (não-IID), que introduzem conflitos entre modelos locais, comprometem a convergência e levam a instabilidade significativa no modelo global [Wang 2024]. Esse fenômeno não apenas degrada o desempenho, mas compromete a confiabilidade e, em muitos casos, inviabiliza o uso prático de FL em sistemas reais.

No domínio de *Learning to Rank* (LTR), uma tarefa central em sistemas de busca, recomendação e apoio à decisão, e ainda um campo emergente no contexto federado, esse problema torna-se ainda mais crítico. Em ambientes federados, os modelos locais capturam padrões altamente específicos de usuários, domínios e contextos, resultando em distribuições fortemente heterogêneas que tornam a agregação substancialmente mais complexa e instável [Yu et al. 2022]. Métodos tradicionais de agregação, como o FedAvg, assumem implicitamente cenários homogêneos e baseiam-se predominantemente no volume de dados, ignorando a qualidade e confiabilidade das atualizações locais — uma limitação crítica em cenários não-IID [McMahan et al. 2023, Li et al. 2020, Reddi and et al. 2021, Wang and Liu 2020].

Em suma, em cenários federados não-IID, contribuições dos clientes devem ser ponderadas por qualidade e estabilidade, não apenas por volume de dados.

Objetivo: Este trabalho tem como objetivo propor uma nova perspectiva de agregação federada orientada por risco e memória temporal, aplicada ao problema de *Federated Learning to Rank* (FLTR) sob cenários não-IID.

A proposta busca melhorar simultaneamente efetividade, estabilidade e consistência do modelo global por meio da integração de dois mecanismos centrais: (i) uma estratégia de agregação sensível ao risco, baseada na variabilidade dos erros de predição dos clientes, e (ii) a incorporação de um mecanismo de memória temporal explícita do modelo global, concebido como princípio de estabilização incremental em aprendizado federado.

Dessa forma, o trabalho busca mitigar os efeitos adversos da heterogeneidade dos dados ao introduzir um mecanismo de agregação mais resiliente e confiável, capaz de equilibrar o trade-off entre privacidade, desempenho e estabilidade. Nesse sentido, contribui para viabilizar o uso prático de modelos de ranqueamento em sistemas distribuídos reais, aproximando o desempenho federado do cenário centralizado mesmo sob restrições de governança e descentralização dos dados.

Contribuição: As principais contribuições desta dissertação são:

- proposição do **FedRisk**, um novo paradigma de agregação federada orientado por qualidade preditiva para FLTR;
- introdução de um mecanismo inédito de ponderação baseado na variabilidade dos erros de predição dos clientes, permitindo diferenciar contribuições locais por confiabilidade;
- proposição de um mecanismo de agregação com memória temporal explícita, no qual o modelo global histórico atua como elemento estruturante de estabilização ao longo do treinamento federado;
- demonstração empírica de que a combinação de sensibilidade ao risco e memória temporal explícita permite simultaneamente melhorar efetividade e reduzir variabilidade em cenários não-IID;
- evidência de que modelos federados de ranqueamento podem se aproximar do desempenho centralizado mesmo sob forte heterogeneidade de dados.

Avanço no estado da arte: Diferentemente de abordagens tradicionais de aprendizado federado — como FedAvg, FedAvgM, FedOpt e FedProx — amplamente utilizadas na literatura, mas limitadas em cenários não-IID [McMahan et al. 2023, Li et al. 2020, Reddi and et al. 2021, Wang and Liu 2020], o método proposto introduz um critério de agregação explicitamente orientado à qualidade preditiva das atualizações locais.

A utilização de sensibilidade ao risco permite reduzir a influência de clientes com comportamento instável ou alta variabilidade de erro, mitigando efeitos negativos da heterogeneidade não-IID. Além disso, a incorporação explícita do modelo global de rodadas anteriores deixa de ser um detalhe metodológico e passa a atuar como um princípio estruturante de estabilização temporal em aprendizado federado, permitindo preservar conhecimento acumulado e reduzir oscilações causadas por atualizações instáveis. A memória temporal deixa de ser um artefato auxiliar e passa a integrar o próprio mecanismo de agregação.

Com isso, a abordagem proposta caracteriza uma nova perspectiva de agregação federada, na qual qualidade preditiva, **estabilidade e memória temporal** são tratadas de forma integrada — um avanço relevante frente às abordagens tradicionais baseadas apenas em **volume de dados** ou **ajustes de otimização**.

Resumo da solução: A solução proposta, denominada **FedRisk**, introduz uma abordagem inédita de agregação federada que combina, de forma integrada, sensibilidade ao risco e memória temporal explícita para lidar com heterogeneidade não-IID em FLTR [Rodrigues et al. 2022, Rodrigues et al. 2025].

A primeira consiste na introdução de um mecanismo de agregação sensível ao risco, no qual as contribuições dos clientes são ponderadas com base na variabilidade de seus erros de predição. Clientes com maior instabilidade ou menor confiabilidade estatística têm sua influência reduzida na agregação global.

A segunda estratégia consiste na incorporação de memória temporal explícita, na qual o modelo global da rodada anterior é combinado com o modelo agregado atual. Essa combinação atua como um mecanismo de estabilização, reduzindo oscilações ao longo do treinamento e preservando conhecimento acumulado em rodadas anteriores.

A integração dessas duas estratégias resulta em um processo de agregação mais robusto, capaz de lidar com heterogeneidade de dados e melhorar a consistência do modelo global em aprendizado federado para ranqueamento.

Avaliação: A avaliação foi realizada utilizando o benchmark MSLR-WEB10K em ambiente federado simulado com 100 clientes, sob distribuições não-IID geradas por particionamento Dirichlet. Os experimentos consideraram 100 rodadas de treinamento e validação cruzada em 5 *folds*.

Os resultados evidenciam ganhos consistentes, expressivos e estatisticamente significativos. O FedRisk alcança um ganho de **+15,6% em nDCG@5** em relação ao FedProx [Li et al. 2020], estabelecendo-se como o melhor desempenho entre os baselines avaliados. Além disso, o método aproxima-se do desempenho centralizado em nDCG@10, evidenciando maturidade metodológica e viabilidade prática da abordagem em cenários reais.

Adicionalmente, observa-se uma **redução consistente da variabilidade entre rodadas e entre diferentes particionamentos dos dados**, refletida na diminuição dos intervalos de confiança entre folds, indicando maior estabilidade do processo de treinamento federado sob heterogeneidade. Esse comportamento evidencia que a abordagem proposta não apenas melhora a efetividade, mas também promove estabilidade consistente, um fator crítico para a viabilidade prática do aprendizado federado em cenários reais.

Análises estatísticas com teste de Wilcoxon e tamanho de efeito confirmam a superioridade prática da abordagem proposta em múltiplas métricas de ranqueamento, reforçando sua robustez em cenários não-IID.

Esses resultados indicam que agregação federada deve evoluir de estratégias baseadas em volume de dados para abordagens orientadas por qualidade, estabilidade e risco, especialmente em cenários não-IID.

Produção científica, técnica e prêmios: Durante o desenvolvimento desta dissertação, foram produzidos trabalhos diretamente derivados da pesquisa principal, bem como colaborações em temas metodologicamente relacionados, incluindo aprendizado sensível ao risco, recuperação de informação, classificação de textos e aplicações de IA.

Produção científica

1. **Aprendizado Federado Incremental e Sensível ao Risco para Modelos de Ranqueamento em Cenários com Distribuições Heterogêneas de Dados** *(autor)*.
   Simpósio Brasileiro de Bancos de Dados (SBBD), 2025, Fortaleza, Brasil.
   Artigo completo diretamente derivado desta dissertação. Recebeu Honra ao Mérito de Melhor Artigo Completo.
2. **Aprendizado Federado Sensível ao Risco em Modelos de Ranqueamento** *(autor)*.
   Anais Estendidos do Simpósio Brasileiro de Bancos de Dados (SBBD), 2023.
   Trabalho precursor que introduziu a linha de pesquisa em FLTR sensível ao risco.
3. **Risk-Sensitive Optimization of Neural Deep Learning Ranking Models with Applications in Ad-Hoc Retrieval and Recommender Systems**.
   Information Processing & Management (Elsevier), 2025 *(coautor)*.
   Trabalho em coautoria. Fundamenta teoricamente o uso de aprendizado sensível ao risco nesta dissertação.
4. **Identificação e Caracterização de Reclamações Duplicadas de Consumidores em Múltiplas Plataformas** *(autor)*.
   Simpósio Brasileiro de Bancos de Dados (SBBD), 2024.
5. **Optimizing Tail-Head Trade-off for Extreme Multi-Label Text Classification (XMTC) with RAG-Labels and a Dynamic Two-Stage Retrieval and Fusion**

Pipeline *(coautor)*.
ACM SIGIR, 2025, Pádua, Itália.
Trabalho em coautoria na área de recuperação híbrida e classificação extrema de textos.

6. **Detecting and Analysing Duplicate Consumer Complaints and Collective Demands Across Multiple Platforms** *(autor)*.
Journal of Information and Data Management (JIDM), 2025.
Artigo aceito. Trabalho em coautoria sobre análise de duplicidade semântica e governança de dados.
7. **Desenvolvimento de um Chatbot Inteligente para Otimização do Suporte aos Usuários dos Sistemas e-SUS APS** *(autor)*.
III Simpósio de Inteligência Artificial em Saúde da UFMG (CIIA Saúde), 2025.
Trabalho apresentado com menção honrosa.

Em conjunto, os resultados desta dissertação estabelecem as bases de uma linha de pesquisa em aprendizado federado sensível ao risco para modelos de ranqueamento sob heterogeneidade de dados.

Prêmios

- **Honra ao Mérito de Melhor Artigo Completo**, SBBD 2025;
- **Menção honrosa**, III Simpósio de Inteligência Artificial em Saúde da UFMG (CIIA Saúde 2025).

Referências

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.

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

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)*.

Reddi, S. J. and et al. (2021). Adaptive federated optimization. In *ICLR*.

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.

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*.

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