Justiça em Modelos de Apoio a Processos de Tomada de Decisão com Uso de Aprendizado Federado

Êrica Peters do Carmo[1], **Agma J. M. Traina**[1], **Caetano Traina Júnior**[1]

[1]Instituo de Ciências Matemáticas e de Computação – Universidade São Paulo (USP)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brazil

ericapetersc@usp.br, agma| caetano@icmc.usp.br

Abstract. *Training imbalanced data can result in biased machine-learning models. In medical decision-making, demographic-based biases can lead to discriminatory treatments for different populations. One possible solution is to increase data diversity through hospital collaboration, but privacy restrictions pose challenges. To deal with this scenario, federated learning allows the development of models using data from multiple institutions while respecting these restrictions. This work proposes using federated learning to develop models with equitable performance across different populations, exploring the potential of this approach in promoting fairness in the medical field.*

Resumo. *Desbalanceamentos nos dados de treinamento podem resultar em modelos de aprendizado de máquina enviesados. Em processos de tomada de decisão na área médica, vieses ligados aos atributos demográficos para diferentes populações podem levar a tratamentos discriminatórios. Uma solução possível é obter dados mais diversos a partir da colaboração entre hospitais, mas restrições de privacidade impõem desafios. Nesse contexto, o aprendizado federado permite desenvolver modelos empregando dados de múltiplas instituições, respeitando essas restrições. Este trabalho propõe o uso do aprendizado federado para desenvolver modelos com desempenho equitativo entre diferentes populações, explorando o potencial dessa abordagem para promover a justiça na área médica.*

1. Introdução

Devido ao seu potencial como ferramenta para a análise de dados e imagens, soluções de inteligência artificial estão se tornando cada vez mais presentes na prática de tomada de decisão na área médica [Al Kuwaiti et al. 2023]. À medida que novas soluções são incorporadas nesta área, surgem desafios quanto à possibilidade de tratamentos discriminatórios e desiguais em relação a diferentes grupos de pacientes [Chen et al. 2023]. Sob a perspectiva do aprendizado de máquina, a disparidade nos resultados dos modelos está relacionada ao conceito de justiça ou equidade (traduzido do termo em inglês "*fairness*").

O desenvolvimento de modelos entendidos como "justos", por sua vez, pode ser comprometido por uma série de vieses provenientes de fontes diversas, incluindo os dados de treinamento, os algoritmos utilizados e até mesmo a sua forma de implementação [Suresh and Guttag 2021]. Especificamente na área da saúde, um possível fator causador de vieses é o desbalanceamento dos dados demográficos dos pacientes. Esse desbalanceamento pode ser explicado pelo fato de que normalmente os dados de cada instituição

de saúde refletem apenas o contexto em que elas estão inseridas e, portanto, apenas as características da população atendida.

Uma solução adotada com frequência para reduzir o desbalanceamento consiste em agregar dados provenientes de múltiplas fontes, obtidos a partir de populações diversas. Contudo, no contexto da saúde, o compartilhamento de dados entre hospitais é altamente regulamentado, devido ao caráter sensível e individual dessas informações. No Brasil, a principal lei que regula o tratamento de dados pessoais, incluindo dados médicos, é a Lei Geral de Proteção de Dados (LGPD) [Brasil 2018].

Nesse cenário, o aprendizado federado oferece uma solução para que instituições colaborem no desenvolvimento de modelos de aprendizado sem comprometer a privacidade dos seus dados. Proposto por McMahan et al. (2017), essa abordagem de aprendizado descentralizado permite criar modelos a partir dos dados de múltiplos clientes. O treinamento ocorre em rodadas onde cada cliente treina um modelo local com seus próprios dados e envia os parâmetros desse modelo ao servidor central, que os agrega para criar um modelo global único. O modelo global é então enviado aos clientes para a próxima rodada de treinamento. Apenas os parâmetros dos modelos são trocados entre servidor e clientes, mantendo os dados individuais locais e privados. Embora o aprendizado federado tenha se tornado popular na área da saúde, seu potencial na redução de vieses demográficos ainda não foi explorado nesse contexto.

Em vista disso, este trabalho propõe explorar essa lacuna de pesquisa, delimitando a seguinte hipótese: *A utilização do aprendizado federado resulta em um modelo de tomada de decisão mais justo do que os modelos treinados de forma centralizada em conjuntos de dados médicos com distribuições distintas de atributos demográficos*? Para direcionar a verificação dessa hipótese, duas questões de pesquisa foram levantadas:

- **RQ1:** Como o desbalanceamento dos atributos demográficos em dados médicos impacta na disparidade de resultados entre grupos populacionais distintos?
- **RQ2:** O aprendizado federado é capaz de gerar um modelo que reduza a disparidade de resultados entre os grupos demográficos quando comparado com os modelos treinados em cada conjunto de dados?

Este trabalho tem como contribuição principal a proposta de uma metodologia experimental para a verificação da hipótese levantada. Nesse sentido, a abordagem proposta possibilita a investigação do potencial do aprendizado federado na redução de vieses originados em dados médicos provenientes de diferentes populações demográficas.

2. Trabalhos Correlatos

Diversos estudos têm investigado como os vieses presentes nos conjuntos de dados médicos impactam os modelos de apoio à tomada de decisão. Por exemplo, Larrazabal et al. (2020) observaram que o desbalanceamento dos dados em relação ao gênero dos pacientes resultou em modelos com desempenho reduzido para o grupo com gênero sub-representado. Em [Seyyed-Kalantari et al. 2020], os autores encontraram disparidades nos resultados de modelos para todos os grupos demográficos, divididos por atributos como sexo, idade e raça.

Quanto ao aprendizado federado, apesar de existirem diversos trabalhos que abordam sua utilização no contexto de tomada de decisão médica, seu potencial na redução de

vieses ainda não foi explorado para esse tipo de problema. Os trabalhos sobre justiça envolvendo grupos demográficos no cenário federado normalmente propõem modificações no método de agregação dos modelos dos clientes [Ezzeldin 2023, Zhang et al. 2020], traçando comparações com o *Federated Averaging* (FedAvg), método comumente utilizado. Além disso, em sua grande maioria, esses trabalhos utilizam conjuntos de dados tabulares. O presente trabalho busca preencher uma lacuna de pesquisa ao explorar a interseção entre justiça e/ou equidade de resultados entre grupos demográficos integrando modelos de tomada de decisão na área médica e aprendizado federado.

3. Metodologia Proposta

A metodologia proposta para verificar a hipótese levantada é um processo experimental, que visa comparar os resultados de modelos treinados de forma centralizada, isolados a cada conjunto de dados, com os resultados do modelo federado. Esse processo experimental é formado por quatro etapas, detalhadas nas próximas seções.

3.1. Seleção dos Dados

A primeira etapa consiste na coleta de diferentes conjuntos de dados médicos a serem utilizados. Este trabalho tem como foco a tarefa de classificação, logo, os conjuntos de dados selecionados devem possuir atributos e rótulos associados à uma mesma tarefa (por exemplo, exames de raio-x torácico para a classificação de doenças pulmonares). Além disso, devem também possuir atributos demográficos relacionados aos pacientes.

A partir dos atributos demográficos comuns aos diferentes conjuntos de dados, os pacientes podem ser divididos em grupos populacionais distintos. Esses grupos podem ser formados tanto em relação a um único atributo (por exemplo, dado o atributo "sexo", é possível dividir os pacientes entre homens e mulheres), quanto a múltiplos atributos (por exemplo, combinação entre atributos "sexo" e "idade"). Nesta etapa, deve-se observar, em cada conjunto de dados, a representatividade dos grupos demográficos formados, buscando identificar vieses que possam comprometer os modelos treinados a partir deles.

3.2. Modelagem Centralizada

A segunda etapa consiste no treinamento de um modelo único para cada conjunto de dados. O termo centralizado está sendo utilizado para se referir a estes modelos por serem treinados de maneira isolada sobre os dados. Além de serem utilizados como *baseline* em relação ao modelo federado, a observação das métricas resultantes desses modelos permite a verificação do impacto dos vieses presentes nos conjuntos de dados.

3.3. Modelagem Federada

A terceira etapa simula um cenário de aprendizado federado onde cada conjunto de dados corresponde a um cliente único. Integrados em um servidor central, os clientes colaboram para que seja gerado um único modelo global. A Figura 1 ilustra as diferenças entre as abordagens centralizada e federada: na primeira, é treinado um modelo exclusivo para cada instituição, usando o seu próprio conjunto de dados; na segunda, esses dados são utilizados no treinamento do modelo local de cada cliente, cujos parâmetros são enviados ao servidor central, que os agrega em um único modelo global.

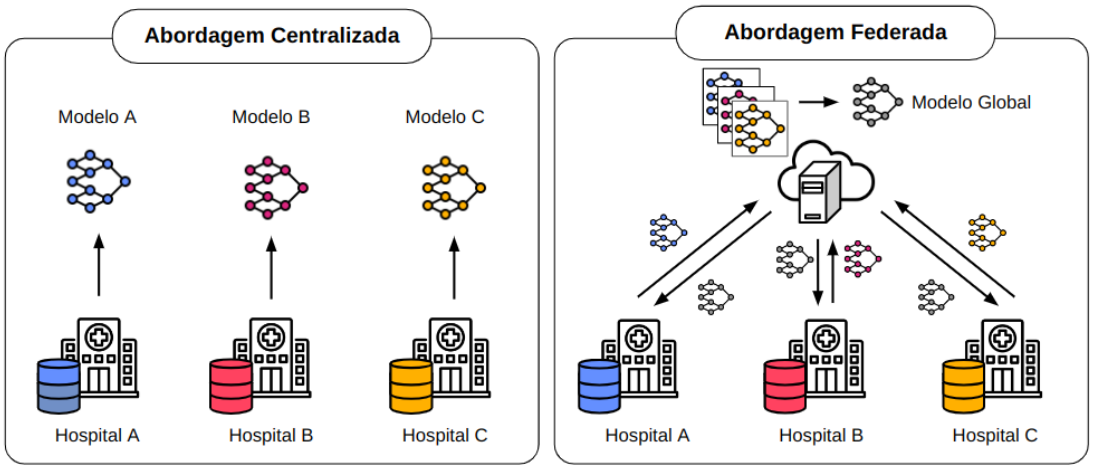

Figura 1. Comparação entre o processo de treinamento dos modelos nas abordagens centralizada e federada

3.4. Avaliação dos Resultados

A última etapa consiste na avaliação dos resultados dos modelos treinados em cada uma das abordagens sob duas perspectivas: justiça (i.e. disparidade de resultados) entre os grupos demográficos e desempenho geral. Na avaliação da justiça, deve-se observar a diferença nas métricas de classificação entre os grupos demográficos de cada conjunto de dados. A comparação entre o resultado dos modelos centralizados e do modelo federado possibilita o entendimento da capacidade de redução de vieses da abordagem federada. A avaliação do desempenho geral, por sua vez, faz-se necessária para garantir que uma possível redução na disparidade dos resultados entre grupos demográficos não seja obtida em detrimento de uma redução significativa no desempenho geral do modelo.

4. Experimento Preliminar

Para avaliar a metodologia proposta, optou-se por utilizar um conjunto de dados reais de fácil compreensão e análise. Desse modo, foi utilizada uma modificação do MNIST [Deng 2012], um conjunto de dados amplamente empregado em tarefas de classificação. A escolha deste *toy dataset* foi feita para dar agilidade e eficiência aos testes iniciais, facilitando a identificação de ajustes necessários antes da realização dos experimentos com dados médicos.

O conjunto de dados MNIST, utilizado para a tarefa de classificação de dígitos manuscritos, é composto por 60.000 dados de treinamento e 10.000 dados de teste, divididos em classes que representam os dígitos de 0 a 9. Como esse conjunto não possui atributos demográficos que possibilitem a divisão entre grupos, foi realizada uma modificação nos dados. Proposta por Arjovsky et al. (2020), a adição de cores aos dígitos permite a separação dos exemplos em grupos distintos. Neste experimento, os exemplos foram divididos em dois grupos conforme o seu rótulo. Aos dígitos menores que quatro foi atribuída a cor vermelha, e aos maiores, a cor verde. A Figura 2 apresenta exemplos dos dados após a adição de cor.

Na primeira etapa do experimento foram criados cinco subconjuntos de dados, cada um representando um cliente distinto. Os dados de treinamento de cada cliente foram

particionados para apresentar desbalanceamento entre os grupos de dígitos, enquanto os dados de teste foram particionados para garantir o balanceamento entre as classes. A distribuição dos grupos nos dados de treinamento para cada cliente pode ser observada na Tabela 1.

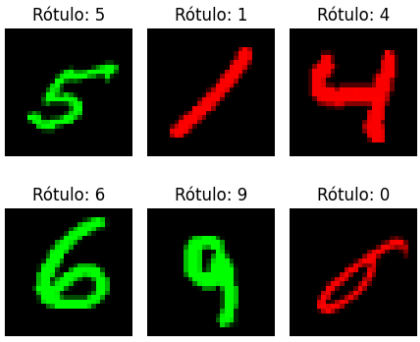

Figura 2. Exemplos do MNIST modificado

Clientes	Nº de Exemplos	Grupo Vermelho	Grupo Verde
Client 0	12806	31%	69%
Cliente 1	13466	52%	48%
Cliente 2	14107	54%	46%
Cliente 3	9250	53%	47%
Cliente 4	10371	69%	31%

Tabela 1. Distribuição dos grupos nos dados de treinamento de cada cliente

Nas duas etapas seguintes, a mesma arquitetura de rede foi empregada para treinar os modelos centralizados e federado. Optou-se por utilizar um rede convolucional composta por duas camadas convolucionais com filtros 5x5 (com 6 e 16 canais, respectivamente), intercaladas com duas camadas de *max pooling* com filtros 2x2. Essas camadas foram seguidas por duas camadas totalmente conectadas com 120 e 84 neurônios, respectivamente, ambas utilizando a função de ativação ReLu. A última camada da rede é totalmente conectada, sendo responsável por produzir as saídas para cada classe.

A otimização do treinamento foi feita com o SGD, utilizando um *momentum* de 0,9 e uma taxa de aprendizado de 0,01. Na modelagem centralizada, cada cliente treinou o seu modelo por 20 épocas, com *batches* de tamanho 10. Já na modelagem federada, o treinamento consistiu em 20 rodadas de comunicação entre o servidor e os clientes. Cada cliente treinou o seu modelo local usando uma época por rodada, também em *batches* de tamanho 10. Para a agregação dos parâmetros dos modelos locais foi utilizado o método FedAvg. Tanto o treinamento centralizado quanto o federado foram executados 5 vezes, mantendo as mesmas partições de dados em todas as execuções.

4.1. Avaliação dos Resultados

De acordo com a metodologia proposta, a avaliação dos resultados considerou tanto a justiça quanto o desempenho geral dos modelos. Para a avaliação da justiça, observou-se a diferença na acurácia dos modelos entre os dois grupos de dígitos. Nota-se que uma menor disparidade nos resultados dos grupos indica uma maior equidade ou justiça do modelo. Conforme é possível observar na Tabela 2, para todos os clientes, os modelos federados resultaram em uma menor disparidade entre as acurácias de ambos os grupos, e portanto maior justiça. Entende-se então que o aumento da diversidade dos dados utilizados no treinamento, possibilitado pela abordagem federada, contribui para a criação de modelos menos enviesados em relação à distribuição dos grupos nos dados de cada cliente.

Clientes	Cliente 0	Cliente 1	Cliente 2	Cliente 3	Cliente 4
Modelo Centralizado	0.023 $\pm$0.018	0.016 $\pm$0.003	0.026 $\pm$0.012	0.081 $\pm$0.053	0.131 $\pm$0.019
Modelo Federado	0.003 $\pm$0.002	0.003 $\pm$0.001	0.007 $\pm$0.003	0.006 $\pm$0.002	0.002 $\pm$0.002

Tabela 2. Média da diferença entre a acurácia dos dois grupos para cinco execuções (valores menores indicam maior justiça)

A avaliação do desempenho geral do modelo é feita, por sua vez, a partir da comparação das acurácias dos modelos centralizados e federado, considerando o conjunto de teste de cada cliente em sua totalidade. É possível observar na Tabela 3 que a acurácia obtida com o modelo federado foi maior do que aquela obtida pelo seu próprio modelo centralizado, para todos os clientes. Este resultado indica um maior potencial de generalização por meio da abordagem federada. Isso é justificado, pois enquanto o modelo centralizado de um cliente possa ser demasiadamente ajustado aos seus dados de treinamento, o modelo federado permite o aprendizado a partir de dados que, apesar de sub-representados em seu conjunto local, estão presentes nos conjuntos dos demais clientes. Além disso, esse resultado garante que a acurácia geral do modelo a ser utilizado pelos clientes não foi comprometida pela diminuição na disparidade entre as acurácias de cada grupo, ou seja, pelo aumento na justiça.

Clientes	Cliente 0	Cliente 1	Cliente 2	Cliente 3	Cliente 4
Modelo Centralizado	0.968 $\pm$0.006	0.975 $\pm$0.003	0.973 $\pm$0.005	0.904 $\pm$0.031	0.919 $\pm$0.004
Modelo Federado	0.995 $\pm$0.001	0.996 $\pm$0.001	0.996 $\pm$0.001	0.996 $\pm$0.001	0.995 $\pm$0.001

Tabela 3. Média da acurácia obtida em cinco execuções para os modelos centralizados e federado

5. Conclusão e Próximos Passos

Este trabalho apresenta uma metodologia para verificar o potencial do aprendizado federado na promoção da justiça em modelos para tomada de decisão. Tratando-se de um trabalho em andamento, foi realizado um experimento preliminar visando validar a metodologia proposta. A partir deste experimento, conclui-se o potencial do aprendizado federado na redução das disparidades de resultados entre grupos, possibilitando o desenvolvimento de modelos mais justos.

Como próximos passos, serão realizados os primeiros experimentos considerando o cenário ideal proposto, formado por três conjuntos de dados médicos voltados à classificação de doenças a partir de radiografias torácicas. Os conjuntos escolhidos são ChestX-ray14 [Wang et al. 2017], CheXpert [Irvin et al. 2019] e MIMIC-CXR [Johnson et al. 2019]. Todos eles possuem informações relacionadas ao sexo e idade dos pacientes, enquanto o último conjunto possui também informações sobre raça e tipo de seguro saúde.

Quanto aos resultados esperados, acredita-se que, por meio da colaboração entre diferentes instituições de saúde, a abordagem federada poderá contribuir para o desenvolvimento de modelos de apoio à tomada de decisão mais equitativos para diferentes grupos demográficos. Se confirmada a hipótese de pesquisa, o aprendizado federado poderá ser entendido como uma alternativa promissora aos métodos tradicionais de mitigação de viés aplicados em conjuntos de dados médicos.

Referências

Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A. V., Al Muhanna, D., and Al-Muhanna, F. A. (2023). A review of the role of artificial intelligence in healthcare. *Journal of personalized medicine*, 13(6):951.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2020). Invariant risk minimization. Preprint disponível em: `https://arxiv.org/abs/1907.02893v3`.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. *Diário Oficial da União*.

Chen, R. J., Wang, J. J., Williamson, D. F., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., and Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Ezzeldin, Y. H. (2023). Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282. PMLR.

Seyyed-Kalantari, L., Liu, G., McDermott, M. B. A., and Ghassemi, M. (2020). Chexclusion: Fairness gaps in deep chest x-ray classifiers. *Pacific Symposium on Biocomputing*, 26:232–243.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Zhang, D. Y., Kou, Z., and Wang, D. (2020). Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060.