Mitigando Impactos de Distribuições Não-IID em Aprendizagem Federada para Sistemas de Recomendação

Arthur Negrão[1], Guilherme C. R. e Rocha[1], Lucas G. dos Santos[1], Pedro I. de S. Malaquias[1], Rodrigo Pedrosa[1], Reinaldo Fortes[1], Pedro Silva[1]

[1]Universidade Federal de Ouro Preto (UFOP)

{arthur.negrao,guilherme.rocha,lucas.gs1,pedro.malaquias}@aluno.ufop.edu.br

{rodrigo.pedrosa,reifortes,silvap}@ufop.edu.br

Abstract. *This work investigates the impact of non-independent and identically distributed data distributions (non-IID) on the performance of movies Recommender Systems (SR) using Federated Learning (FL). Through the collaboration between parts, FL allows the distribution of computational cost and expands the data diversity available to the training process, favoring the construction of SRs with good generalization capacity. In this context, and to analyze the effectiveness of the partial data sharing strategy, experiments were conducted in four scenarios using the dataset Movielens1M: (C1) IID data; (C2) non-sharing non-IID data; (C3) non-IID data with 1% sharing; and (C4) non-IID data with 5% sharing. Two federated paradigms were used: cross-silo and cross-device. Experimental results indicate that partial data sharing is a promising approach to mitigate the adverse effects of non-IID distributions in federated learning, incurring approximately 9,5% increases for NDCG and 0.046 drops in the MAE, for example. Thus, predictive performance, privacy, and computational cost are balanced.*

Resumo. *Este trabalho investiga o impacto de distribuições de dados não-Independente e Identicamente Distribuída (não-IID) no desempenho de Sistemas de Recomendação (SR) de filmes utilizando Aprendizagem Federada, em inglês Federated Learning (FL). Através da colaboração entre partes, o FL permite a distribuição do custo computacional e amplia a diversidade de dados disponíveis para o processo de treinamento, favorecendo a construção de SRs com boa capacidade de generalização. Neste contexto, e com o objetivo de analisar a eficácia da estratégia de compartilhamento parcial de dados, conduziram-se experimentos em quatro cenários utilizando o dataset MovieLens1M: (C1) dados IID; (C2) dados não-IID sem compartilhamento; (C3) dados não-IID com 1% de compartilhamento; e (C4) dados não-IID com 5% de compartilhamento. Dois paradigmas federados foram utilizados: Cross-Silo e Cross-Device. Resultados experimentais indicam que o compartilhamento parcial de dados é uma abordagem promissora para mitigar os efeitos adversos de distribuições não-IID em Aprendizagem Federada, incorrendo em aumentos de aproximadamente 9,5% para NDCG e quedas de 0,046 no MAE, por exemplo. Assim sendo, equilibra-se desempenho preditivo, privacidade e custo computacional.*

1. Introdução

O mundo contemporâneo caracteriza-se pela constante digitalização das mais diversas relações e sistemas. Sejam compras, comunicação, lazer, finanças ou até mesmo contratos formais: quase tudo é feito em frente a um computador através da internet [Lü et al. 2012].

Consequentemente, tal fenômeno incorre na constante e massiva geração de registros eletrônicos por parte de tais aplicações, contribuindo diretamente com a realidade vigente de *big data* [Jagadish 2015] e a problemática do paradoxo da escolha. De maneira simples, essa pode ser explicada como a dificuldade humana de tomar uma decisão quando uma quantidade vasta de possibilidades faz-se presente [Fernandez 2017]. Tomando como exemplo a indústria do cinema, os usuários das plataformas de *streaming* podem ter dificuldades em escolher um filme diante da vastidão de opções do catálogo, constituindo um problema tanto para os mesmos quanto para os *stakeholders* do ramo. Isso porque estes anseiam o melhor desfruto possível das plataformas por parte dos seus usuários, tendo em vista que a satisfação destes tem como consequência geral seu consumo reiterado [Sharma and Gera 2013].

Mediante tal problemática surgem os Sistemas de Recomendação (SR), os quais fazem uso de métodos computacionais para aprender perfis de usuário a fim de realizar recomendações personalizadas. Para tal, os mesmos necessitam de uma quantidade considerável de dados e recursos computacionais para que seu treinamento resulte em um modelo eficiente e acurado [Sharma and Gera 2013] - e, muitas vezes, estes podem ser de difícil obtenção. Na direção de mitigar tal problema, a Aprendizagem Federada (em inglês *Federated Learning* - FL) apresenta-se como um promissor agente à medida que permite a colaboração de instituições/empresas na construção e treinamento de modelos preditivos eficazes através de uma estrutura de rede. Nesta estrutura, cada nó é responsável pelo treinamento de um modelo local que será enviado a um servidor orquestrador, o qual, por sua vez, agregará os modelos recebidos de acordo com uma regra pré-estabelecida [Kairouz et al. 2021].

Em [Kairouz et al. 2021], em termos de FL, explicam que as redes projetadas assumem natureza *Cross-Silo* (CS) ou *Cross-Device* (CD). O primeiro geralmente é associado às redes de treinamento entre instituições (*e.g.* hospitais, universidades e/ou empresas) e seus *datacenters*, onde cada nó da rede apresenta capacidades computacionais consideráveis e usualmente totalizam algo na faixa de 2-100 clientes na rede. Já o ambiente CD caracteriza-se pela presença massiva de clientes (até 10^{10}), onde cada cliente pode ser um dispositivo móvel, um agente *IoT* ou um computador pessoal, entre outros. Neste cenário, a confiabilidade, disponibilidade e capacidade computacional de cada cliente são muito menores, mas em troca a rede ganha uma dispersão considerável - o que pode favorecer à medida que os dados se originam de fontes muito mais diversas.

Entretanto, é comum que, em cenários reais de aplicação da estratégia federada, estabeleça-se uma distribuição enviesada dos dados entre os nós da rede - o que é conhecido na literatura como não-Independente e Identicamente Distribuída (não-IID). Conforme explicam [Zhu et al. 2021], as distribuições não-IID podem caracterizar-se de diversas formas, como por exemplo através do desbalanceamento na proporção de instâncias por classe para cada nó (podendo haver o desconhecimento de uma ou mais classes em um determinado nó), ou por meio da extrema disparidade entre as classes e os nós, onde as *features* de uma classe ocorrem somente em um nó. Ainda sobre o tema, os autores explicam que tais distribuições dificultam o processo de convergência de treinamento dos modelos, haja vista que contribuem para a divergência dos modelos locais e, consequentemente, para a queda de desempenho do processo de agregação dos modelos.

O objetivo principal deste artigo é avaliar a eficiência da estratégia proposta em [Zhao et al. 2018] no contexto de SRs federados. A escolha da estratégia foi baseada

na eficiência da mesma reportada pelos autores, que muitas vezes equiparava-se ou superava outras estratégias da literatura, em associação com sua versatilidade e facilidade de implantação - o que permite sua adequação a *pipelines* federados das mais diversas naturezas. Vale relembrar que a estratégia referida consiste no compartilhamento parcial dos dados entre os nós da rede, no intuito de mitigar os efeitos negativos causados pela realidade não-IID das distribuições. Adicionalmente, o trabalho também busca mensurar a queda de desempenho provocada pela distribuição não-IID em relação a uma distribuição IID. Para auxílio no cumprimento destes objetivos, estabeleceram-se algumas Perguntas de Pesquisa (PP), que contribuíram como um norte para o desenvolvimento e realização dos experimentos, tal qual a análise de seus resultados. Elas são apresentadas a seguir:

PP1: A estratégia de compartilhamento é capaz de solucionar ou mitigar a problemática causada pela natureza não-IID dos dados?

PP2: Para as métricas avaliativas estabelecidas, em quais faixas estabelecem-se as quedas esperadas pela transição de uma distribuição IID para uma não-IID?

Mediante tais objetivos, este trabalho conduziu experimentos no sentido de avaliar o processo de treinamento federado sob âmbito não-IID no contexto de SRs cinematográficos. Mais especificamente, foi mensurada e avaliada a queda de desempenho gerada pelas distribuições não-IID mediante à tarefa de recomendação proposta no conjunto de dados *MovieLens1M* em quatro contextos diferentes: (C1) dados IID; (C2) dados não-IID sem compartilhamento; (C3) dados não-IID com 1% de compartilhamento dos dados entre os nós; e (C4) dados não-IID com 5% de compartilhamento dos dados entre os nós. Ademais, também avaliaram-se os impactos da estratégia neste contexto supracitado para com o desempenho dos modelos gerados em dois paradigmas de FL diferentes: *Cross-Silo* e *Cross-Device*.

Os resultados indicaram que a estratégia de compartilhamento parcial foi eficaz em mitigar as quedas de desempenho geradas pelas distribuições não-IID. Constataram-se aumentos de 9,5% para NDCG@10 e quedas de 0,046 no MAE mediante à aplicação da estratégia em relação à sua não aplicação. Além disso, a estratégia de compartilhamento parcial foi comparada com a *FedProx* - uma conhecida estratégia da literatura voltada para mitigação de danos causados pelos dados não-IID - e provou-se mais eficiente que a mesma para a tarefa proposta.

Este trabalho é organizado da seguinte forma: na Seção 2 é feita uma revisão bibliográfica acerca da literatura relativa ao tema deste artigo; na Seção 3 apresenta-se a metodologia proposta para este trabalho; na Seção 4 são apresentados os resultados obtidos através dos experimentos propostos, além de uma discussão acerca dos mesmos; e, por fim, são apresentadas as conclusões e trabalhos futuros na Seção 5.

2. Revisão Bibliográfica

Esta seção apresenta uma revisão da literatura acerca do tema deste trabalho. Em mais detalhes, a Seção 2.1 discorre sobre SRs de maneira geral e a Seção 2.2 sobre aprendizagem federada e dados não-IID.

2.1. Sistemas de Recomendação

A literatura científica no ramo de SRs voltados a filmes/cinema é extremamente rica e volumosa. Pesquisas nesta área buscam, de maneira geral, desenvolver métodos capazes de recomendar obras cinematográficas a diferentes usuários de maneira personalizada, *i.e.* levando em consideração suas preferências pessoais. Em [Ahuja et al. 2019], por exemplo, os autores apresentam uma proposta de recomendação de filmes baseada nos métodos de

k-médias (em inglês *k-means*) e k-vizinhos mais próximos (em inglês *K-Nearest Neighbors* - KNN). De maneira sucinta, tais métodos podem ser caracterizados pelo objetivo de particionar um conjunto de dados em k subconjuntos de tal forma que cada instância i seja atribuída ao subconjunto w que apresente as instâncias com maior similaridade a i - havendo um critério pré-estabelecido para mensurar tal similaridade. Em termos performáticos, utilizando a *Root Mean Squared Error* (RMSE) como métrica de avaliação, a proposta atingiu o valor de 1,08 no *dataset MovieLens-100k*.

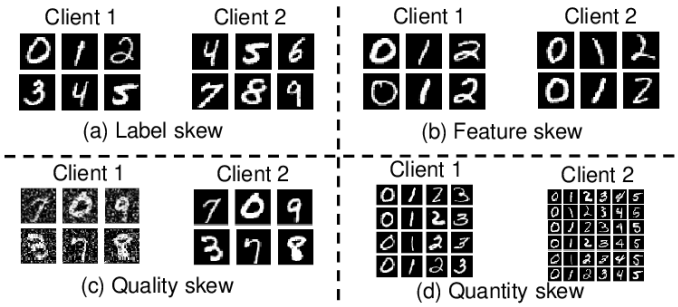
Já em [Christakou et al. 2007] é apresentada outra proposta de recomendação filmográfica baseada em filtragem híbrida através de redes neurais. Os autores explicam que as mesmas foram treinadas para prover uma estimativa da avaliação do usuário sobre o filme, sendo tal avaliação expressa como um valor discreto $r \in [1, K]$. E, através de um limiar $\omega \in [1, K]$, tais predições são classificadas entre "filme recomendado" ou "filme não recomendado" para o respectivo usuário. Com base neste processo de agregação booleana, foram conduzidos experimentos utilizando o *dataset MovieLens-100k* com $K = 5$ e $\omega = 3$, e obteve-se um resultado de 82% de recomendações adequadas. Ademais, a obra também apresenta uma interessante investigação acerca da lógica *fuzzy*, onde experimentou-se com operadores da mesma e obtiveram-se taxas de precisão na faixa de 85%.

Em [Shahbazi and Byun 2019] apresenta-se um sistema de recomendação de produtos baseado em *XGBoost* (acrônimo para *Extreme Gradient Boosting*), a qual consiste em uma técnica distribuída e escalável para construção de árvores de decisão cuja construção é apoiada pelo algoritmo de descida de gradiente [Li et al. 2019]. Os autores utilizaram uma base de dados baseada em compras *online* da cidade de Jeju, na Coreia do Sul, e obtiveram resultados na faixa de 89,6% de acurácia.

2.2. Aprendizagem Federada e Dados Não-IID

Uma distribuição Independente e Identicamente Distribuída (IID) é toda aquela em que cada amostra tem a mesma probabilidade de ocorrer e não influencia em amostras futuras [Arafeh et al. 2022]. Portanto, por definição, uma distribuição Não-IID não respeita tais condições, caracterizando-se pelo enviesamento das amostras. A Figura 1 apresenta alguns possíveis exemplos de como as amostras podem ser enviesadas.

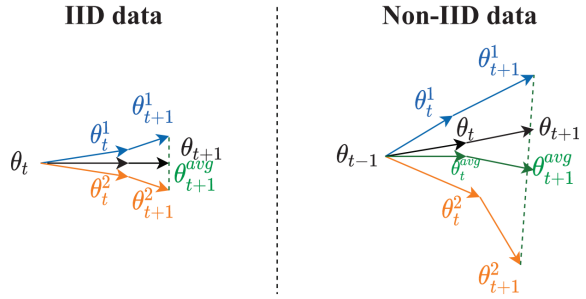
Figura 1. Exemplos de distribuições não-IID. Veja que, em *a*, cada cliente não conhece todas as classes possíveis. Em b, a natureza/característica dos dados entre amostras difere consideravelmente. Em *c*, o mesmo ocorre com a qualidade dos dados. E em *d* há uma desproporção de quantidade de instâncias por cliente. Fonte: [Zhu et al. 2024].



Um levantamento da literatura científica é feito em [Zhu et al. 2021] acerca do tema de aprendizagem federada com distribuições de dados não-IID. Explica-se na obra que o treinamento das redes federadas com tal condição normalmente acompanha uma

notável degradação da performance obtida. Tal fato pode ser explicado pela divergência dos modelos locais gerados na rede que, quando agregados, não são capazes de caminhar em prol de uma representação eficaz dos dados, inviabilizando a convergência do modelo. A Figura 2 ilustra tal processo.

Figura 2. Divergência de modelos gerada pelos dados não-IID. Fonte: [Zhu et al. 2021].



[Zhao et al. 2018] apresenta uma estratégia para mitigar o problema de distribuições não-IID em redes federadas. Ela consiste no compartilhamento de dados entre os nós da rede de forma a gerar-se um pequeno *subset* de dados global que estará disponível a todos os clientes da rede em seu processo de treinamento. Para que a estratégia seja eficaz, é importante ressaltar que o processo de construção de tal *subset* deve almejar a construção de uma distribuição IID. Em termos de eficiência, os autores conduziram experimentos com a base de dados CIFAR-10 e reportaram um incremento na faixa de 30% de acurácia com o compartilhamento de apenas 5% do *dataset*.

Ainda sobre tal estratégia, entende-se que o compartilhamento dos dados fere um dos princípios motivadores das redes federadas: a privacidade. Assim sendo, a depender do cenário de aplicação, é extremamente importante considerar a viabilidade desta troca de privacidade por desempenho. Por outro lado, é importante mencionar que propostas como a criptografia homomórfica [Zhang et al. 2020, Fang and Qian 2021], adição de ruído nos dados e seleção de *features* não sensíveis para compartilhamento podem ser capazes de mitigar - ou até mesmo solucionar - tais problemas de privacidade.

Transitando para a obra de [Ali et al. 2021], os autores apresentam uma arquitetura federada (Fed-CARS) associada a um protocolo de privacidade (UDCP) que permite aos usuários definir o "grau de colaboração" dos mesmos para com a rede federada. Mais detalhadamente, este permite que usuários definam políticas para o compartilhamento de seus dados privados contextuais para com seus respectivos modelos locais. Em experimentos realizados com *MovieLens100k*, a estratégia dos autores atingiu valores como 0,79 para MAE, 0,96 para RMSE e 0,85 para NDCG. Já [Ammad-Ud-Din et al. 2019] apresentam um SR construído sobre arquitetura federada para recomendação de filmes especificamente através de filtragem colaborativa, *i.e.* filtragem baseada em avaliações de outros usuários sobre um mesmo item. Os autores atingiram RMSE de 0,7, precisão de 0,3, e *recall* de 0,13 no *dataset MovieLens1M*.

Em [Huang et al. 2022] é realizada uma revisão sobre o FL em ambiente *Cross-Silo*. Os autores explicam que, neste cenário, o número de clientes é pequeno e espera-se que cada cliente participe ativamente no processo de treinamento durante toda a sua extensão. Os autores também explicam que, em CS, espera-se que cada cliente detenha uma quantidade considerável de dados - e não uma pequena porção, como ocorre em CD;

e que tarefas como predição de doenças ou cálculo de risco de seguros são bons exemplos de aplicações do paradigma CS.

Já o paradigma *Cross-Device* caracteriza-se pela vastidão de clientes distribuídos nos mais diversos dispositivos. De maneira contrária ao CS, é comum que um cliente não contribua ativamente durante a integridade do processo de treinamento, estabelecendo-se menor disponibilidade e confiabilidade de cada cliente [Karimireddy et al. 2021, ur Rehman et al. 2021]. Ademais, também é esperado que cada cliente possua um conjunto de dados menor e mais restrito quando comparado ao paradigma CS, o que reforça e aumenta a probabilidade da ocorrência de distribuições não-IID em ambiente CD [Lin et al. 2022].

Por fim, a pesquisa sobre a literatura revelou que, apesar de existirem muitos trabalhos na literatura propondo novas técnicas/estratégias para mitigar os impactos gerados pelos dados não-IID sobre treinamentos federados, poucos são os trabalhos que avaliam de maneira sistemática e criteriosa tais estratégias em contextos específicos. Assim sendo, pontua-se como uma importante contribuição deste trabalho à literatura científica a avaliação empírica da estratégia de [Zhao et al. 2018] sobre o contexto cinematográfico.

3. Metodologia

Esta seção apresenta a metodologia utilizada para a execução deste trabalho. A Seção 3.1 apresenta o *dataset* utilizado e a manipulação de dados realizada sobre o mesmo; a Seção 3.3 apresenta o processo de avaliação dos modelos treinados; a Seção 3.2 apresenta o modelo utilizado para treinamento; e a Seção 4.1 apresenta a configuração experimental. O código-fonte deste trabalho pode ser encontrado na plataforma *GitHub* [1].

3.1. Base de Dados e Pré-processamento

A base de dados utilizada foi a *MovieLens1M*[2] em sua versão de fevereiro de 2003, cuja escolha, ao invés da *MovieLens32M*, foi baseada na necessidade de encontrar um balanço entre complexidade dos experimentos e viabilidade computacional. Ela é composta por 1.000.210 avaliações de 3.954 filmes distintos realizadas por 6.040 usuários únicos, onde cada avaliação única apresenta uma nota (*rating*) de 0 a 5 (valores inteiros) que o usuário deu ao filme em questão. O pré-processamento dos dados contou com a normalização Min-Máx de todos os atributos numéricos; a seleção das *features movieId*, *userId*, *timestamp*, *age* e *genres*; e o processo de *one-hot encode* do atributo *genres*. A Figura 3 apresenta um histograma sobre a quantidade de avaliações realizadas por cada indivíduo.

Para execução dos experimentos, o *dataset* utilizado foi dividido em 70% das instâncias para treino e 30% para testes. Para o particionamento não-IID utilizaram-se o Particionador Patológico [Li et al. 2022, Beutel et al. 2020] com atribuições de classe realizadas no modo primeiro determinístico baseadas no rótulo *rating* e limitadas por um valor máximo de $M = 2$ distintas notas por partição; e o Particionador de Dirichlet [Beutel et al. 2020] com $\alpha = 0,05$ (quanto menor o valor mais não-IID são as partições geradas [Jimenez et al. 2024]) baseado no atributo *userId*. Este foi selecionado para simular o que potencialmente ocorre no mundo real, onde as diferentes plataformas de *streaming* conhecem apenas uma parcela dos indivíduos que consumem conteúdo cinematográfico. Nestes cenários, as partições geradas podem apresentar *label skew*, *quantity skew* ou ambos (Figura 1). Já para o particionamento IID utilizou-se uma implementação convencional [Beutel et al. 2020]. A Figura 4 ilustra o funcionamento destes.

[1]Disponível em `https://github.com/arthurnfmc/NonIID-MovieRecommendation/`

[2]Disponível em `https://grouplens.org/datasets/movielens/1m/`

Figura 3. Número de avaliações por usuário distinto.

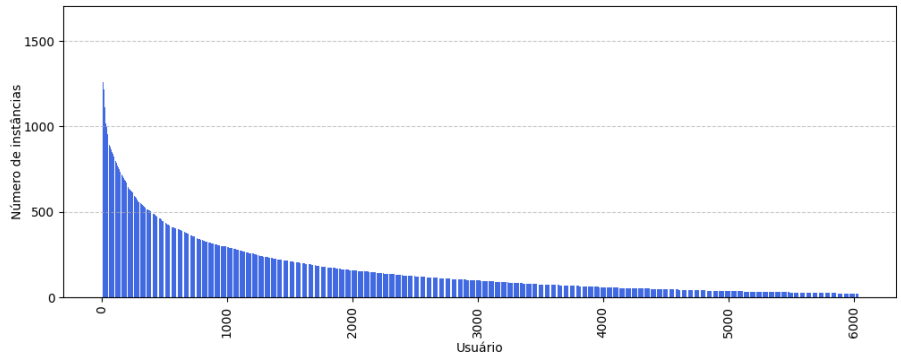
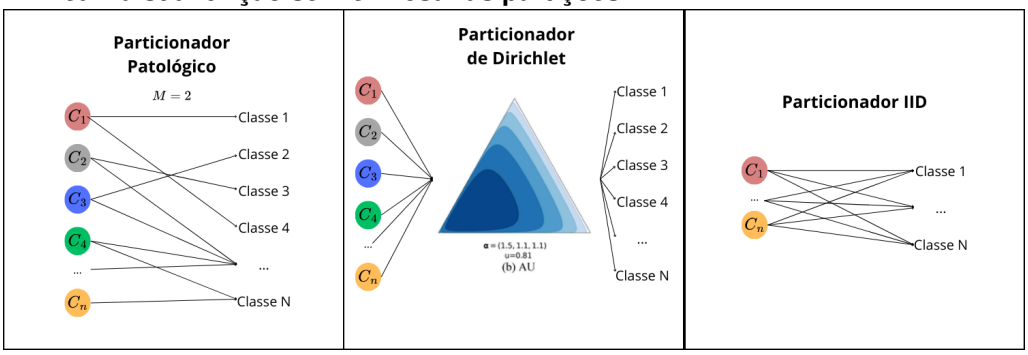


Figura 4. O Particionador Patológico permite que cada cliente conheça no máximo M classes, de forma que toda classe é conhecida por pelo menos um cliente. O Particionador de Dirichlet respeita a distribuição de mesmo nome para construir as partições enviesadas conforme o parâmetro α [Li et al. 2023a]. O Particionador IID realiza sua função sem enviesar as partições.



3.2. Modelo Utilizado

Como o enfoque deste artigo é avaliar a eficiência da estratégia proposta em [Zhao et al. 2018] no contexto de SR para responder as perguntas de pesquisa propostas nestes trabalho, e não necessariamente obter o melhor desempenho possível em termos das métricas de avaliação, optou-se por uma arquitetura de rede neural simples - haja visto o sucesso das redes neurais em outros trabalhos da literatura [Li et al. 2023b, Zhang et al. 2018, Zhang 2022] e a economia de recursos/tempo computacional ao adotar-se uma rede simples. Vale destacar que o próprio trabalho de [Zhao et al. 2018] utiliza uma CNN simples (2 camadas convolucionais seguidas de 2 densas) na tarefa de avaliar sua proposta de compartilhamento no *dataset* CIFAR-10.

Em mais detalhes, a arquitetura da rede implementada neste trabalho é sequencialmente descrita por: **1.** uma camada de entrada com 22 neurônios; **2.** uma camada densa com 64 neurônios e ativação ReLU; **3.** uma camada densa com 32 neurônios e ativação ReLU; **4.** uma camada de saída com 1 neurônio. E sobre sua configuração para o treinamento, foi empregado o otimizador Adam e a função de perda adotada foi o MSE.

3.3. Processo de Avaliação

O processo de avaliação contou com as métricas do Erro Médio Absoluto (MAE, *Mean Absolute Error* em inglês), Raiz do Erro Quadrado Médio (RMSE, *Root Mean Squared Error* em inglês), Precisão@10 (Pre@10), Recall@10 (Rec@10) e NDCG@10 (acrônimo de *Normalized Discounted Cumulative Gain*). As duas primeiras representam a diferença entre os valores preditos e os valores reais, respectivamente, sem nenhuma forma de

ponderação e com maior punição para erros crassos em relação a pequenos erros. Ou seja, quanto menor o valor obtido em tais métricas, melhor é o SR em realizar predições.

Já as outras três métricas amparam-se sobre o conceito de relevância. Para este trabalho, uma recomendação considerada relevante é aquela cuja nota dada ao filme é maior que 3; haja visto que as notas 1 e 2 são consideradas ruins, a nota 3 é considerada neutra e as notas 4 e 5 são consideradas boas. De tal maneira, Pre@10 e Rec@10 mensuram, respectivamente, a porcentagem das recomendações relevantes dentre as *top*-10 recomendações e o número de recomendações relevantes nas *top*-10 sobre o total de recomendações relevantes - e, assim sendo, o melhor valor que podem atingir é 1 (100%). Vale lembrar que, quando o número total de recomendações relevantes é muito maior que 10, é de se esperar que Rec@10 assuma valores baixos. Por fim, o NDCG@10 o quão boa é a ordem dos *top*-10 filmes recomendados, levando em consideração sua relevância e sua posição na lista de recomendações. O melhor valor que esta métrica pode atingir é 1 (100%).

Com base em tais métricas, os modelos treinados durante os experimentos foram avaliados conforme a diretiva federada através de uma média ponderada pelo número de instâncias de cada cliente. Ou seja, o valor reportado para uma métrica arbitrária do modelo global corresponde à média dos valores obtidos por cada cliente, o qual possui um conjunto de teste individual que pondera tal valor de acordo com seu tamanho.

Por fim, para comparação estatística entre os valores encontrados experimentalmente, adotou-se o teste t de Welch como teste comparativo, assumindo **1.** distribuição normal e que **2.** os desvios padrão das duas populações não são necessariamente iguais. O teste foi configurado no formato bi-caudal com 95% de significância, onde $H_0 : \mu_1 = \mu_2$ e $H_1 : \mu_1 \neq \mu_2$.

4. Experimentos e Resultados

Esta seção apresenta os experimentos propostos e seus respectivos resultados. Em mais detalhes, a Seção 4.1 descreve o processo experimental; a Seção 4.2 apresenta os resultados deste processo; e a Seção 4.3 promove uma discussão sobre os resultados encontrados.

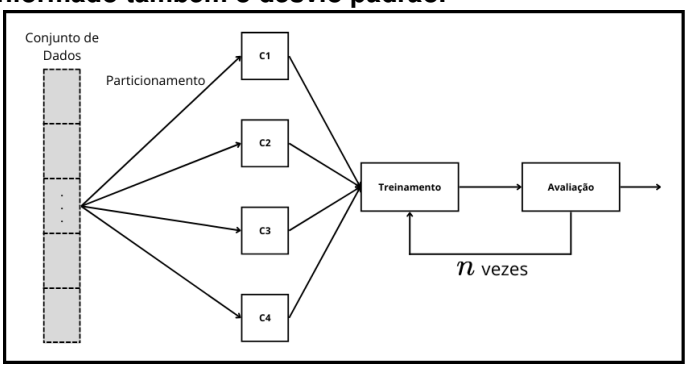
4.1. Configuração dos Experimentos

A experimentação consistiu no treinamento federado do modelo descrito previamente no intuito de avaliar a eficácia da estratégia de compartilhamento de dados de [Zhao et al. 2018]. O processo de agregação de modelos adotado foi o *Federated Averaging* (FedAvg) [McMahan et al. 2016], e os mesmos foram treinados em quatro diferentes cenários: particionando os dados de treino de forma IID (C1); particionando os dados de treino de forma não-IID sem compartilhar dados conforme a estratégia supracitada (C2); particionando os dados de treino de forma não-IID compartilhando 1% dos mesmos conforme a estratégia supracitada (C3); particionando os dados de treino de forma não-IID compartilhando 5% dos mesmos conforme a estratégia supracitada (C4). O valor de 5% foi escolhido seguindo o trabalho de [Zhao et al. 2018], enquanto o valor de 1% foi escolhido para analisar a eficiência da estratégia mediante situações com maior escassez de dados compartilhados. Também é importante ressaltar que C1 funcionará como um valor de referência comparativa, e em nenhum momento emprega distribuições não-IID ou mesmo a estratégia de compartilhamento em seu *pipeline*.

A experimentação avaliou as diferenças nos resultados entre os ambientes *cross-silo* e *cross-device*. E para efeitos de robustez e comparação estatisticamente fundamentada dos resultados, cada configuração experimental foi executada $n = 33$ vezes. O resultado é

expresso através da média de seus resultados e seu desvio padrão. A Figura 5 resume o processo descrito.

Figura 5. *Pipeline* do processo experimental. Inicialmente, os dados são particionados conforme o cenário experimental. Após tal, ocorre o treinamento e avaliação do modelo n vezes. O resultado é reportado como a média dos resultados, sendo informado também o desvio padrão.



Os cenários supracitados foram simulados em ambiente CS e CD, onde o primeiro contou com 20 clientes na rede federada e o segundo com 200. Infelizmente, devido a indisponibilidade de mais recursos computacionais, não foi possível a experimentação com uma quantidade maior de clientes. Em ambos os casos, realizaram-se dois grupos de experimentos: um utilizando o Particionador Patológico (PP) e outro utilizando o Particionador de Dirichlet (PD), ambos em conformidade com a descrição da Seção 3.1. Com tais particionadores, para fins de comparação, também foram executadas instâncias experimentais utilizando a *FedProx* [Li et al. 2020], outra estratégia para mitigar o impacto dos dados não-IID em treinamentos federados. Em termos sucintos, seu funcionamento consiste na adição de um termo punitivo na função de perda que cresce à medida que o modelo local diverge do modelo global. Por fim, sobre a configuração de hiper-parâmetros, cada execução constou com 3 épocas globais (*i.e.* rodadas de agregação), 5 épocas locais (*i.e.* épocas de treinamento em cada cliente local) e *learning rate* $\alpha = 0,001$. Os mesmos foram selecionados após testes empíricos que consistiram em três execuções do *pipeline* com um número reduzido de clientes (cinco no total), onde dois hiper-parâmetros eram fixados e o outro era apresentado à variações. Ao final das execuções, era escolhido aquele que apresentasse o menor RMSE.

Os experimentos foram implementados utilizando o *framework Flower* [Beutel et al. 2020] em sua versão 1.18 através de seu módulo de simulações. Os mesmos foram conduzidos em ambiente *Docker* (v. 27.3.1) operando sobre dois computadores: o primeiro contando com uma CPU AMD Ryzen Threadripper 3960X com 24 cores físicos (48 threads) a 3.70GHz e 128GB RAM DDR4, e o segundo contando com um Intel i9-10900 com 10 cores físicos (20 threads) de 2.80GHz e 128GB RAM DDR4.

4.2. Resultados Experimentais

Os experimentos foram conduzidos conforme a descrição da seção anterior, sendo seus resultados expressos nas tabelas 1 e 2 no formato $\mu \pm \sigma$.

Iniciando a análise com os resultados da experimentação CD, é possível afirmar que, para todas as métricas de ambos os particionadores, com exceção da Pre@10 utilizando o PP, a estratégia de compartilhamento parcial em seu grau mais intenso (5%) obteve o melhor resultado. Para NDCG@10, por exemplo, o incremento de C4 em relação a

Tabela 1. Resultados experimentais para 20 clientes (*Cross-Silo*). Em negrito o melhor resultado para cada métrica. Aproximações até a terceira casa decimal.

Casos/Métricas	C1 (IID)	C2 (Não-IID)	C3 (1%)	C4 (5%)	*FedProx*
Particionador Patológico (PP)					
MAE	**0,867 ± 0,003**	0,911 ± 0,004	0,888 ± 0,003	0,869 ± 0,003	0,910 ± 0,004
RMSE	1,063 ± 0,002	1,087 ± 0,003	1,073 ± 0,002	**1,062 ± 0,002**	1,086 ± 0,002
Pre@10	0,593 ± 0,114	**0,598 ± 0,117**	**0,598 ± 0,117**	**0,598 ± 0,118**	**0,598 ± 0,117**
Rec@10	**0,001 ± 0,000**	**0,001 ± 0,000**	**0,001 ± 0,000**	**0,001 ± 0,000**	**0,001 ± 0,000**
NDCG@10	**0,880 ± 0,031**	0,857 ± 0,036	0,873 ± 0,031	0,876 ± 0,036	0,869 ± 0,031
Particionador de Dirichlet (PD)					
MAE	0,868 ± 0,003	0,869 ± 0,003	0,866 ± 0,002	**0,861 ± 0,002**	0,868 ± 0,003
RMSE	1,063 ± 0,001	1,064 ± 0,002	1,062 ± 0,002	**1,058 ± 0,001**	1,063 ± 0,002
Pre@10	**0,598 ± 0,117**	**0,598 ± 0,117**	**0,598 ± 0,117**	**0,598 ± 0,117**	**0,598 ± 0,117**
Rec@10	**0,001 ± 0,000**	**0,001 ± 0,000**	**0,001 ± 0,000**	**0,001 ± 0,000**	**0,001 ± 0,000**
NDCG@10	0,873 ± 0,040	**0,880 ± 0,021**	0,867 ± 0,033	0,868 ± 0,034	0,869 ± 0,036

Tabela 2. Resultados experimentais para 200 clientes (*Cross-Device*). Em negrito o melhor resultado para cada métrica. Aproximações até a terceira casa decimal.

Casos/Métricas	C1 (IID)	C2 (Não-IID)	C3 (1%)	C4 (5%)	*FedProx*
Particionador Patológico (PP)					
MAE	0,888 ± 0,003	0,912 ± 0,012	0,881 ± 0,002	**0,866 ± 0,002**	0,912 ± 0,002
RMSE	1,084 ± 0,002	1,094 ± 0,008	1,079 ± 0,002	**1,065 ± 0,002**	1,093 ± 0,002
Pre@10	**0,581 ± 0,039**	0,567 ± 0,057	0,572 ± 0,032	0,574 ± 0,030	0,574 ± 0,030
Rec@10	**0,007 ± 0,000**	**0,007 ± 0,001**	**0,007 ± 0,000**	**0,007 ± 0,000**	**0,007 ± 0,000**
NDCG@10	0,793 ± 0,027	0,767 ± 0,024	0,823 ± 0,021	**0,862 ± 0,013**	0,765 ± 0,026
Particionador de Dirichlet (PD)					
MAE	0,887 ± 0,002	0,887 ± 0,002	0,879 ± 0,002	**0,866 ± 0,002**	0,887 ± 0,002
RMSE	1,084 ± 0,002	1,084 ± 0,001	1,078 ± 0,001	**1,065 ± 0,001**	1,083 ± 0,001
Pre@10	**0,574 ± 0,030**	**0,574 ± 0,030**	**0,574 ± 0,030**	**0,574 ± 0,030**	**0,574 ± 0,030**
Rec@10	**0,007 ± 0,000**	**0,007 ± 0,000**	**0,007 ± 0,000**	**0,007 ± 0,000**	**0,007 ± 0,000**
NDCG@10	0,788 ± 0,022	0,792 ± 0,018	0,829 ± 0,022	**0,863 ± 0,011**	0,794 ± 0,029

C1 é de aproximadamente 7%, e a queda do MAE é de mais de 0,02 no mesmo cenário comparativo. Ou seja, a estratégia não só é capaz de mitigar a problemática não-IID, como também supera - com significância estatística - os resultados obtidos em distribuições IID. Já para CS tais incrementos de performance são mais amenos, numa relação em que C4 mais se aproxima de C1 do que necessariamente o supera. Um reflexo disto é que, para a métrica MAE, por exemplo, a diferença entre C1 e C4 é de 0,007 e -0,002 respectivamente utilizando o PD e o PP.

Realocando o enfoque da análise sobre os diferentes particionadores, é possível afirmar que os mesmos impactam diretamente no desempenho do modelo treinado. Em início, tratando do PD, o qual constrói as partições com base no atributo identificador do usuário, não observaram-se diferenças estatisticamente significantes entre C1 e C2 para nenhuma métrica. Uma hipótese válida para justificar tal fato é que, tanto em CS quanto CD, construíram-se partições com quantidades razoáveis de usuários do *dataset* que, como demonstram os resultados, foram suficientes para gerar uma boa representação dos dados. Por outro lado, o mesmo não ocorre com o PP: como pode ser visto pelas métricas MAE e NDCG@10, quedas de desempenho estatisticamente significantes ocorrem entre C1 e C2. Em mais detalhes, há um aumento de 0,044 no MAE e uma queda de 2,3% no NDCG@10 entre os respectivos cenários. Isso indica que cada cliente da rede conhecer filmes com no máximo $M = 2$ notas distintas é um empecilho considerável no processo de treinamento federado e, consequentemente, dificulta a construção de uma boa representação dos dados.

Ainda sobre os dados das tabelas 1 e 2, percebe-se que a estratégia de compartilhamento parcial obteve melhor desempenho geral que a *FedProx* no *dataset* avaliado. Tal diferença é especialmente notória na configuração CD com o PP, onde entre C4 e *FedProx* podem ser observados, por exemplo, aumentos de 0,03 na RMSE e quedas de 10% na

NDCG@10, aproximadamente. Os mesmos dados também revelam que os ganhos de performance do modelo são diretamente proporcionais à quantidade de dados compartilhados - evidenciando, portanto, o *trade-off* existente entre desempenho e privacidade dos dados.

Por fim, um ponto de importante menção é a similaridade geral entre todos os resultados obtidos para Pre@10 e Rec@10. Isso é um indicativo de que, indiferentemente do cenário experimental e da presença (ou não) da estratégia, as top-10 recomendações do sistema, ao menos neste *dataset*, tendem a permanecer iguais - caracterizando filmes potencialmente muito populares. Apesar de ser uma forma segura de realizar recomendações (já que escolhas populares usualmente agradam diversos públicos), tal realidade pode prejudicar a visibilidade de novos filmes e a inovação nas recomendações.

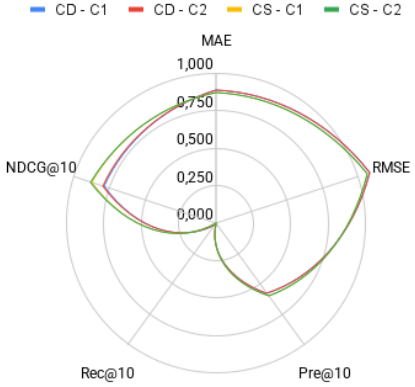
4.3. Discussão dos Resultados

Realizadas as ponderações sobre os dados obtidos, segue-se com a análise das perguntas de pesquisa estabelecidas anteriormente. Iniciando com a PP1, os resultados obtidos indicam que a estratégia de compartilhamento é de fato capaz de mitigar a problemática causada pela natureza não-IID dos dados e, em alguns casos, até mesmo superar os resultados obtidos em partições IID. Como pode ser visto nas tabelas 1 e 2, há uma melhora de desempenho diretamente proporcional à quantidade de dados compartilhados nos treinamentos com distribuições não-IID. Ademais, é possível afirmar que, pelo menos nas configurações estabelecidas neste trabalho, a estratégia de compartilhamento parcial apresenta desempenho melhor que a *FedProx*, uma conhecida estratégia da literatura científica neste ramo.

Prosseguindo para a PP2, observou-se nos experimentos com o PP que os dados não-IID dificultaram a construção de uma boa representação dos dados. Em termos numéricos, tal dificuldade refletiu, por exemplo, em quedas de 2,3% no NDCG@10 e aumentos de 0,044 no MAE entre C1 e C2 em ambiente CS. Portanto, tendo em vista as regras de particionamento do PP descritas na Seção 3.1, é razoável afirmar que, para treinamento eficaz do SR, não basta a um cliente da rede federada conhecer avaliações de filmes restritas a um conjunto pequeno ($M = 2$) de notas; mas sim que é necessário conhecer avaliações dispersas em um amplo conjunto de diferentes *ratings*.

Entretanto, de maneira inesperada, o mesmo fenômeno não ocorreu para os experimentos com o PD. Isto é, não houve queda de desempenho estatisticamente considerável entre C1 e C2 tanto em CD quanto em CS, a ver pela Figura 6.

Figura 6. Comparativo IID vs. Não-IID em ambiente CD e CS para o Particionador de Dirichlet. Veja que, para cada ambiente, as linhas se sobrepõem quase perfeitamente.



Tendo em vista que o PD baseia-se no atributo de *userId*, interpreta-se tal fato como um indicativo de que, ao menos para o problema de recomendação de filmes, não é

uma necessidade irremediável conhecer todos os usuários que realizam avaliações sobre os filmes do *dataset* e nem mesmo que estes organizem-se em proporções uniformes; mas sim que é suficiente conhecer uma porção razoável dos mesmos para ser apto a construir uma boa representação dos dados.

5. Conclusões e Trabalhos Futuros

Este trabalho conduziu e avaliou o treinamento de uma rede federada com dados não-IID aplicando o compartilhamento parcial de dados entre clientes, proposto por [Zhao et al. 2018], no intuito de mitigar os problemas gerados pela natureza não-IID dos dados sobre a tarefa de recomendação cinematográfica. Esta tarefa consiste em um problema de regressão cujo objetivo é prever a nota dada por um determinado usuário a um filme. A experimentação foi conduzida em quatro diferentes cenários, tanto em ambiente CS quanto CD.

Os resultados experimentais indicaram que a estratégia de compartilhamento parcial dos dados é eficaz em mitigar a problemática gerada pelas distribuições não-IID. Observaram-se, por exemplo, reduções de 0,046 do MAE e aumentos de 9,5% do NDCG@10 com o compartilhamento de 5% dos dados (C4) em ambiente CD com o PP em relação à não utilização da estratégia (C2). Adicionalmente, observou-se que a estratégia de compartilhamento parcial apresentou melhor desempenho que a *FedProx* na tarefa proposta, a ver pelos dados das tabelas 1 e 2.

Ainda sobre os resultados, observa-se que a experimentação com o PD apresentou pouca ou nenhuma diferença entre os cenários IID e não-IID - a ver pela Figura 6. Tal fato evidencia que, ao menos para a tarefa deste trabalho, é suficiente que cada nó da rede federada conheça uma porção razoável de usuários para a construção de uma boa representação dos dados. Em outras palavras: não é uma necessidade absoluta conhecer todos os usuários para ser capaz de realizar boas recomendações de maneira geral.

Apesar de revelar fatos interessantes, pontua-se que esta pesquisa possui algumas limitações, como, por exemplo, a quantidade reduzida de clientes em CD e a falta da análise dos efeitos da estratégia de compartilhamento mediante técnicas criptográficas/de segurança de dados. Desta maneira, para trabalhos futuros, a experimentação com maior quantidade de clientes no ambiente CD poderia revelar novas dinâmicas da tarefa analisada, bem como a utilização de criptografia homomórfica para aumentar a segurança dos dados, por exemplo. Também seria interessante a condução de experimentos que realizassem o particionamento com base em atributos diferentes (e.g. o gênero do filme avaliado), ou mesmo com outros particionadores e/ou bases de dados mais robustas. Ademais, a substituição da rede neural como modelo utilizado pelo XGBoost poderia constituir uma análise de performance instigante, já que tal técnica é reconhecida pelo alto desempenho sobre dados tabulares [Shwartz-Ziv and Armon 2022]; ou então a utilização de um modelo *transformer-based*, *e.g.* BERT4Rec. Finalmente, explorar outros domínios de aplicação, para além dos filmes, seria interessante para comprovar a capacidade de generalização dos métodos.

Agradecimentos

Os autores agradecem ao apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, projeto APQ-01768-24), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Universidade Federal de Ouro Preto (PROPPI/UFOP).

Referências

[Ahuja et al. 2019] Ahuja, R., Solanki, A., and Nayyar, A. (2019). Movie recommender system using k-means clustering and k-nearest neighbor. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 263–268. IEEE.

[Ali et al. 2021] Ali, W., Kumar, R., Deng, Z., Wang, Y., and Shao, J. (2021). A federated learning approach for privacy protection in context-aware recommender systems. *The Computer Journal*, 64(7):1016–1027.

[Ammad-Ud-Din et al. 2019] Ammad-Ud-Din, M., Ivannikova, E., Khan, S. A., Oyomno, W., Fu, Q., Tan, K. E., and Flanagan, A. (2019). Federated collaborative filtering for privacy-preserving personalized recommendation system. *arXiv preprint arXiv:1901.09888*.

[Arafeh et al. 2022] Arafeh, M., Hammoud, A., Otrok, H., Mourad, A., Talhi, C., and Dziong, Z. (2022). Independent and identically distributed (iid) data assessment in federated learning. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 293–298. IEEE.

[Beutel et al. 2020] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., and Lane, N. D. (2020). Flower: A friendly federated learning research framework. *CoRR*, abs/2007.14390.

[Christakou et al. 2007] Christakou, C., Vrettos, S., and Stafylopatis, A. (2007). A hybrid movie recommender system based on neural networks. *International Journal on Artificial Intelligence Tools*, 16(05):771–792.

[Fang and Qian 2021] Fang, H. and Qian, Q. (2021). Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet*, 13(4):94.

[Fernandez 2017] Fernandez, C. (2017). The paradox of choice: why more is less. *Vikalpa*, 42(4):265–267.

[Huang et al. 2022] Huang, C., Huang, J., and Liu, X. (2022). Cross-silo federated learning: Challenges and opportunities. *arXiv preprint arXiv:2206.12949*.

[Jagadish 2015] Jagadish, H. V. (2015). Big data and science: Myths and reality. *Big Data Research*, 2(2):49–52.

[Jimenez et al. 2024] Jimenez, G. D. M., Anagnostopoulos, A., Chatzigiannakis, I., and Vitaletti, A. (2024). Fedartml: A tool to facilitate the generation of non-iid datasets in a controlled way to support federated learning research. *IEEE Access*.

[Kairouz et al. 2021] 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 machine learning*, 14(1–2):1–210.

[Karimireddy et al. 2021] Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S., Stich, S. U., and Suresh, A. T. (2021). Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676.

[Li et al. 2022] Li, Q., Diao, Y., Chen, Q., and He, B. (2022). Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE.

[Li et al. 2020] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

[Li et al. 2019] Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction based on xgboost algorithm. *Frontiers in genetics*, 10:1077.

[Li et al. 2023a] Li, X., Fei, J., Xie, J., Li, D., Jiang, H., Wang, R., and Qi, Z. (2023a). Open set recognition for malware traffic via predictive uncertainty. *Electronics*, 12(2):323.

[Li et al. 2023b] Li, X., Sun, L., Ling, M., and Peng, Y. (2023b). A survey of graph neural network based recommendation in social networks. *Neurocomputing*, 549:126441.

[Lin et al. 2022] Lin, D., Guo, Y., Sun, H., and Chen, Y. (2022). Fedcluster: A federated learning framework for cross-device private ecg classification. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE.

[Lü et al. 2012] Lü, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., and Zhou, T. (2012). Recommender systems. *Physics reports*, 519(1):1–49.

[McMahan et al. 2016] McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. (2016). Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629.

[Shahbazi and Byun 2019] Shahbazi, Z. and Byun, Y.-C. (2019). Product recommendation based on content-based filtering using xgboost classifier. *Int. J. Adv. Sci. Technol*, 29:6979–6988.

[Sharma and Gera 2013] Sharma, L. and Gera, A. (2013). A survey of recommendation system: Research challenges. *International Journal of Engineering Trends and Technology (IJETT)*, 4(5):1989–1992.

[Shwartz-Ziv and Armon 2022] Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

[ur Rehman et al. 2021] ur Rehman, M. H., Dirir, A. M., Salah, K., Damiani, E., and Svetinovic, D. (2021). Trustfed: A framework for fair and trustworthy cross-device federated learning in iiot. *IEEE Transactions on Industrial Informatics*, 17(12):8485–8494.

[Zhang et al. 2020] Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., and Liu, Y. (2020). {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning. In *2020 USENIX annual technical conference (USENIX ATC 20)*, pages 493–506.

[Zhang et al. 2018] Zhang, L., Luo, T., Zhang, F., and Wu, Y. (2018). A recommendation model based on deep neural network. *IEEE Access*, 6:9454–9463.

[Zhang 2022] Zhang, Y. (2022). Music recommendation system and recommendation model based on convolutional neural network. *Mobile Information Systems*, 2022(1):3387598.

[Zhao et al. 2018] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

[Zhu et al. 2021] Zhu, H., Xu, J., Liu, S., and Jin, Y. (2021). Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390.

[Zhu et al. 2024] Zhu, S., Zeng, J., Wang, S., Sun, Y., Li, X., Yao, Y., and Peng, Z. (2024). On admm in heterogeneous federated learning: Personalization, robustness, and fairness. *arXiv preprint arXiv:2407.16397*.