Esquecer é Preciso: Um Estudo sobre o Impacto da Remoção de Dados no Desaprendizado de Máquinas

Milena Curtinhas Santos[1], **João Paulo de Brito Gonçalves**[2]
Antonio A. de A. Rocha[3], **Rodolfo da Silva Villaça**[1]

[1]Departamento de Informática (DI/CT)
Universidade Federal do Espírito Santo (Ufes)
milena.c.santos@edu.ufes.br, rodolfo.villaca@ufes.br

[2]Instituto Federal do Espírito Santo (Ifes)
Campus Cachoeiro do Itapemirim/ES
jpaulo@ifes.edu.br

[3]Instituto de Computação (IC)
Universidade Federal Fluminense (UFF)
arocha@ic.uff.br

Abstract. *The increasing stringency of data protection regulations, such as the LGPD and GDPR, has driven the development of machine unlearning techniques to ensure the right to be forgotten in artificial intelligence models. This article reviews key concepts, challenges, and recent advances in the field, experimentally evaluating different unlearning algorithms, including DaRE and DynFrs, across multiple datasets. Results indicate that small-scale data removals generally have a limited impact on model accuracy, highlighting the need for efficient and robust approaches. Finally, future perspectives are discussed, such as validating unlearning through blockchain and integrating explainable AI (XAI) techniques, aiming for more transparent and trustworthy systems.*

Resumo. *O crescente rigor das legislações de proteção de dados, como a LGPD e o GDPR, impulsionou o desenvolvimento de técnicas de desaprendizado de máquina (machine unlearning) para garantir o direito ao esquecimento em modelos de inteligência artificial. Este artigo revisa conceitos, desafios e avanços recentes na área, avaliando experimentalmente diferentes algoritmos de desaprendizado, como DaRE e DynFrs, em múltiplos conjuntos de dados. Os resultados mostram que pequenas remoções de dados tendem a ter impacto limitado na acurácia dos modelos, mas ressaltam a importância de abordagens eficientes e robustas. Por fim, são discutidas perspectivas futuras, incluindo a validação do desaprendizado via blockchain e a integração com técnicas de IA explicável (XAI), visando sistemas mais transparentes e confiáveis.*

1. Introdução

Na atual "era dos dados", a prática do esquecimento intencional de informações pessoais tem ganhado destaque, especialmente com a promulgação da Lei Geral de Proteção de Dados (LGPD) [Presidência da República 2018]. Nesse contexto, a retenção de dados pode infringir diretrizes de privacidade, sobretudo diante da coleta massiva de dados pessoais. O ato de desaprender, definido como o processo de perda ou esquecimento daquilo que foi

aprendido, tem sido alvo de estudos recentes para garantir a privacidade dos indivíduos. Além disso, regulamentações como o Regulamento Geral de Proteção de Dados (RGPD) europeu [União Europeia 2016] garantem o "direito ao esquecimento" [Dang 2021], permitindo que indivíduos solicitem a remoção de seus dados.

A coleta de informações pessoais é fundamental para o desenvolvimento de modelos de aprendizado de máquina em diversas aplicações. Esses dados permitem que esses algoritmos identifiquem padrões de comportamento e são extremamente valiosos para grandes corporações. Contudo, a segurança associada ao acesso a essas informações pessoais tem gerado preocupações, pois pode representar uma ameaça à privacidade dos usuários. Embora a remoção de dados das bases de dados *backend* atenda à regulamentação, isso não é suficiente no contexto da inteligência artificial. Modelos de aprendizado de máquina são criados a partir da compressão dos dados de treinamento, e alguns são profundamente adaptados ao seu treino. Além disso, modelos mais complexos, como os de aprendizado profundo, dificultam a identificação da ligação entre os dados e os parâmetros do modelo. Retreinar modelos para excluir dados solicitados muitas vezes não é uma opção, pois esse processo é computacionalmente dispendioso e representa um entrave significativo de desempenho. Portanto, é evidente o desafio de estabelecer uma estratégia que permita aos modelos de aprendizado de máquina "esquecer" informações previamente aprendidas a partir dos dados de treinamento.

O desaprendizado de máquinas parte do pressuposto de que o modelo já "aprendeu". Para suportar os pedidos de esquecimento torna-se necessário o provisionamento de um mecanismo de desaprendizado. Por definição, esse mecanismo recebe como entrada um conjunto de dados de treino, um conjunto de dados a serem esquecidos e um modelo de aprendizado. O resultado será um modelo "desaprendido", que tem como hipótese final ser um modelo que se aproxima do modelo retreinado a partir do zero, sem os dados removidos. Portanto, o problema central do desaprendizado de máquina envolve a comparação entre duas distribuições de modelos de aprendizado de máquina: o desaprendido e o retreinado [Bourtoule et al. 2021, Brophy and Lowd 2021, Thudi et al. 2022].

O objetivo deste artigo é investigar o comportamento de dois algoritmos de desaprendizado de máquina, DaRe [Brophy and Lowd 2021] e DynFrs [Wang et al. 2025], em diferentes cenários de remoção de dados e sob distintas distribuições dos conjuntos de dados utilizados nos experimentos. A proposta contempla uma avaliação empírica do impacto da remoção de amostras tanto em contextos com dados independentemente e identicamente distribuídos (IID) quanto em cenários non-IID, utilizando múltiplos *datasets* amplamente adotados na literatura.

Nos experimentos conduzidos para avaliar estratégias de desaprendizado de máquina, inicialmente foram analisados diferentes classificadores, Random Forest, Decision Tree e Extreme Gradient Boosting, utilizando o conjunto de dados MNIST [Alpaydin and Kaynak 1998]. Em uma etapa subsequente, os algoritmos DaRe e DynFrs, foram selecionados para avaliação da acurácia dos modelos frente a diferentes taxas de remoção de dados, comparando os resultados obtidos com o retreinamento tradicional dos algoritmos. Os experimentos foram realizados utilizando os conjuntos de dados Vaccine, Adult, Bank, Diabetes, NoShow, Synthetic e MNIST. A análise busca quantificar os efeitos do desaprendizado sobre a acurácia dos modelos resultantes, contribuindo para a compreensão da robustez e eficiência dessas abordagens.

2. Trabalhos Relacionados

Diversas abordagens têm sido propostas para endereçar o problema do desaprendizado. O método mais direto é o retreinamento completo do modelo a partir dos dados remanescentes após a exclusão dos exemplos a serem esquecidos. No entanto, essa abordagem é computacionalmente custosa, especialmente para modelos complexos e grandes conjuntos de dados, tornando-se impraticável em diversos cenários [Bourtoule et al. 2021].

No contexto de modelos baseados em árvores e florestas aleatórias, destacam-se trabalhos como DaRE e DynFrs. O DaRE [Brophy and Lowd 2021] propõe um método para remover a influência de exemplos específicos em florestas aleatórias, garantindo que o modelo resultante seja indistinguível de um modelo treinado sem os dados removidos. O trabalho demonstra que é possível realizar o desaprendizado de maneira eficiente, utilizando estratégias de particionamento e atualização seletiva dos nós das árvores. Em sequência, o DynFrs [Wang et al. 2025] amplia essas ideias, apresentando um framework que permite a remoção dinâmica de dados em florestas aleatórias, com ganhos significativos de eficiência computacional em comparação ao retreinamento completo.

Outra linha de investigação explora métodos para garantir o desaprendizado em ambientes federados, onde os dados são distribuídos entre múltiplos clientes. F2L2 [Jin et al. 2024] aborda o problema do desaprendizado certificado em modelos lineares de forma federada. O método proposto oferece garantias formais de que o modelo resultante não retém qualquer influência dos dados removidos. Ainda no contexto de modelos baseados em árvores, o HedgeCut [Schelter et al. 2021] introduz uma abordagem para manter conjuntos de árvores aleatórias sob a remoção de pequenas frações de dados de treinamento. No domínio de sistemas de recomendação, o UltraRE [Li et al. 2023] aprimora o *framework RecEraser*, que utiliza um esquema baseado em *ensemble* para garantir o desaprendizado completo. O método é validado em conjuntos de dados reais, demonstrando superioridade em relação a abordagens anteriores em termos de desempenho e eficiência.

Diante da diversidade de abordagens e dos avanços recentes no campo do desaprendizado de máquina, torna-se fundamental realizar uma avaliação criteriosa e comparativa dessas técnicas. Considerando os diferentes contextos de aplicação, este artigo se propõe a analisar de forma sistemática o desempenho e a eficácia dos algoritmos de desaprendizado DaRe e DynFrs, escolhidos pela disponibilidade de código e reprodutibilidade dos experimentos já publicados. O objetivo é identificar estratégias que possibilitem a remoção eficiente e segura de informações nestes algoritmos, preservando tanto a privacidade quanto a utilidade dos modelos, e fornecendo subsídios para a escolha de métodos adequados em cenários práticos.

3. Desaprendizado de Máquinas

O conceito de desaprendizado, ou *machine unlearning*, surge como resposta à necessidade de garantir, em sistemas de aprendizado de máquina, o cumprimento do "direito ao esquecimento" previsto em regulamentações como o GDPR, que exige que dados pessoais possam ser removidos a pedido do usuário [Nguyen et al. 2024, Bourtoule et al. 2021]. No entanto, a simples exclusão dos dados não é suficiente quando modelos de aprendizado de máquina já foram treinados, pois esses modelos podem ter memorizado padrões ou informações específicas dos dados excluídos, tornando possível, por meio de ataques

de privacidade, inferir ou recuperar informações individuais. Assim, o desaprendizado se torna um desafio, com uma questão fundamental: como garantir que um modelo de aprendizado de máquina, após a remoção de um subconjunto específico de dados, não retenha qualquer influência desses dados em seu comportamento futuro?

Formalmente, o desaprendizado pode ser definido como o processo pelo qual um modelo treinado em um conjunto de dados é ajustado de modo a remover completamente a influência de um subconjunto específico desses dados, de tal forma que o modelo resultante seja indistinguível de um modelo que nunca tenha sido treinado com os dados excluídos. Em outras palavras, dada uma função de aprendizado $A(D)$, que produz um modelo a partir de um conjunto de dados D, e um subconjunto D_f a ser esquecido, o desaprendizado exige que o modelo obtido após o desaprendizado, $U(D, D_f, A(D))$, siga a mesma distribuição de probabilidade que o modelo treinado diretamente no conjunto restante, $A(D \setminus D_f)$ [Nguyen et al. 2024]. Essa exigência de indistinguibilidade é essencial para garantir que o modelo não carregue traços dos dados removidos.

A principal diferença entre desaprendizado e retreinamento está na eficiência e na garantia técnica oferecida [Bourtoule et al. 2021]. O retreinamento consiste em simplesmente treinar um novo modelo a partir do zero, utilizando apenas os dados remanescentes após a exclusão dos dados a serem esquecidos. No entanto, o retreinamento pode ser extremamente custoso em termos computacionais e de tempo, especialmente para modelos complexos e grandes conjuntos de dados, pois exige que todo o processo de aprendizagem seja repetido sempre que uma exclusão é solicitada. Em contraste, o desaprendizado busca alternativas mais eficientes para alcançar resultados similares, sem a necessidade de retreinamento completo.

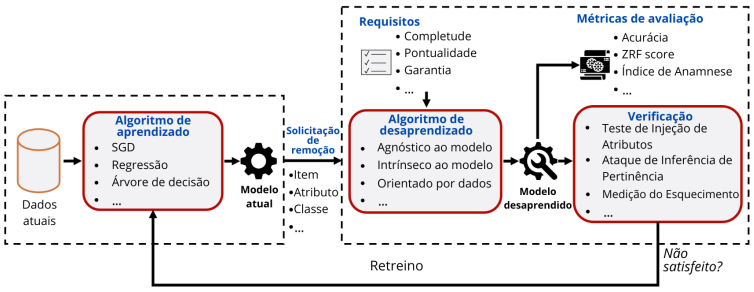


Figura 1. Fluxograma do desaprendizado. Adaptado de: [Nguyen et al. 2024].

O desafio do desaprendizado é agravado pela natureza estocástica e incremental dos processos de treinamento [Nguyen et al. 2024]. Modelos complexos, como redes neurais profundas, são treinados em lotes aleatórios de dados, e a ordem dos lotes pode variar entre diferentes épocas de treinamento, tornando difícil rastrear a influência exata de cada ponto de dados nos parâmetros do modelo. Além disso, o treinamento é incremental, o que significa que cada atualização do modelo reflete todas as atualizações anteriores, tornando a remoção seletiva de influências ainda mais complexa [Bourtoule et al. 2021]. A Figura 1 apresenta um fluxograma detalhado do processo de desaprendizado, mostrando como o modelo é treinado, submetido a solicitações de exclusão, e verificado quanto à privacidade e eficácia da remoção.

4. Metodologia de Experimentação

A abordagem metodológica deste estudo foi estruturada em duas etapas complementares e sequenciais, visando uma análise abrangente sobre os efeitos da remoção de dados em modelos de aprendizado de máquina. No **Experimento 1** buscou-se investigar como diferentes estratégias de remoção de instâncias impactam a acurácia das predições dos modelos, estabelecendo um referencial comparativo para as etapas subsequentes. No **Experimento 2**, a atenção foi direcionada à avaliação do desaprendizado propriamente dito, com ênfase na comparação entre os resultados obtidos por algoritmos específicos de desaprendizado e aqueles provenientes do retreinamento tradicional dos modelos.

4.1. Conjuntos de Dados

A escolha dos datasets usados em cada etapa da experimentação foi conduzida conforme seus objetivos específicos, detalhados a seguir. A Tabela 1 sumariza as principais características dos conjuntos utilizados.

- **Experimento 1:** Utilizou-se o conjunto de dados MNIST [1], referência clássica para tarefas de classificação. O MNIST é composto por 70.000 imagens de dígitos (60.000 para treino e 10.000 para teste), cada uma representada por uma matriz de 28x28 pixels.
- **Experimento 2:** Foram empregados múltiplos conjuntos de dados, abrangendo diferentes domínios e características, a saber: Vaccine [2], Adult [3], Bank [4], Diabetes [5], NoShow [6], Synthetic [7] e MNIST [8]. Esses conjuntos apresentam variações quanto ao número de instâncias, atributos, proporção de classes e presença de variáveis categóricas.

Tabela 1. Resumo dos conjuntos de dados utilizados nos experimentos.

Dataset	# Instâncias	# Atributos	Proporção Classe (+)
MNIST	70.000	784	10 classes
Vaccine	26.707	185	46,4%
Adult	48.842	107	23,9%
Bank	41.188	63	11,3%
Diabetes	101.766	253	46,1%
NoShow	110.527	99	20,2%
Synthetic	1.000.000	40	50,0%

Os experimentos realizados confirmaram a eficácia dos algoritmos DaRE e DynFrs em diferentes domínios, demonstrando desempenho consistente em bases tabulares e no conjunto MNIST. Os resultados evidenciam que, no cenário avaliado, tais métodos

[1] https://www.kaggle.com/datasets/hojjatk/mnist-dataset
[2] https://www.drivendata.org/competitions/66/flu-shot-learning/data/
[3] https://doi.org/10.24432/C5XW20
[4] https://doi.org/10.24432/C5K306
[5] https://doi.org/10.24432/C5230J
[6] https://www.kaggle.com/datasets/joniarroba/noshowappointments
[7] https://anonymous.4open.science/r/DynFrs-2603/README.md
[8] https://www.kaggle.com/datasets/hojjatk/mnist-dataset

são capazes de executar desaprendizado seletivo de forma eficiente, mantendo níveis satisfatórios de acurácia preditiva mesmo após a remoção de amostras. Essas evidências reforçam o potencial de aplicação prática dos algoritmos em cenários que exigem conformidade com legislações de proteção de dados, como LGPD e GDPR, assegurando o direito ao esquecimento sem comprometer significativamente a utilidade dos modelos de aprendizado de máquina.

4.2. Modelos e Algoritmos Avaliados

A metodologia de experimentação usada neste trabalho foi conduzida em duas etapas distintas, com o objetivo de avaliar os impactos do desaprendizado de máquina na acurácia de modelos de classificação.

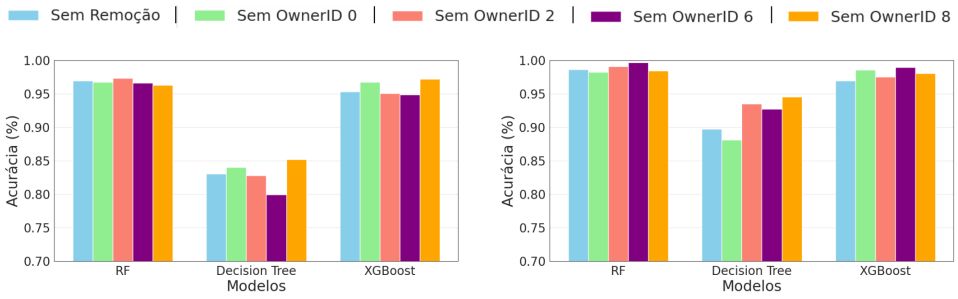
No **Experimento 1**, foram utilizados algoritmos tradicionais de aprendizado de máquina supervisionado para tarefas de classificação, tais como: Random Forest, Decision Tree, Extreme Gradient Boosting (XGBoost). O treinamento foi realizado sobre diferentes conjuntos de dados, citados anteriormente. Após o treinamento inicial, foram realizadas remoções pontuais de exemplos nos dados de entrada, seguidas do retreinamento completo dos modelos a partir dos dados remanescentes. Essa etapa visou estabelecer uma linha de base para mensurar o efeito direto da exclusão de amostras sobre a acurácia final dos modelos.

No **Experimento 2**, os algoritmos DaRE e DynFrs, projetados especificamente para o cenário de desaprendizado, foram empregados em diferentes contextos de remoção de dados. Esses contextos consideraram tanto cenários com dados independentemente e identicamente distribuídos (IID) quanto cenários com distribuições non-IID, de modo a avaliar a robustez das abordagens em situações mais realistas e desafiadoras. Em ambas as etapas experimentais, a métrica principal de avaliação foi a acurácia final dos modelos resultantes após o processo de remoção, com o objetivo de comparar a eficácia e o impacto das diferentes estratégias de desaprendizado adotadas.

4.3. Procedimentos Experimentais

No **Experimento 1**, o objetivo foi analisar o efeito da exclusão de instâncias em modelos clássicos de classificação, de modo a estabelecer um referencial (baseline) para comparação posterior com algoritmos projetados especificamente para desaprendizado. Para isso, foram utilizados modelos treinados com o conjunto de dados MNIST, amplamente aceito na literatura devido à sua padronização e caráter representativo. Após o treinamento inicial, diferentes proporções de instâncias foram removidas do conjunto de treino. As remoções foram conduzidas sob duas estratégias distintas: uma com amostras independentemente e identicamente distribuídas (IID) e outra com dados non-IID, simulando cenários mais realistas de exclusão de dados. A acurácia dos modelos foi registrada após cada cenário de remoção, possibilitando a quantificação precisa do impacto das exclusões sobre o desempenho preditivo.

O **Experimento 2** teve como foco a avaliação comparativa dos algoritmos DaRE e DynFrs, ambos voltados para a realização de desaprendizado em modelos baseados em florestas de decisão. Esses algoritmos foram aplicados a diferentes conjuntos de dados, incluindo Vaccine, Adult, Bank, Diabetes, NoShow, Synthetic e MNIST, abrangendo uma diversidade de domínios e tamanhos de dados. Os parâmetros de ambos os métodos foram



(a) non-IID **(b)** IID

Figura 2. Gráficos comparativos da remoção de *ownerID* em diferentes divisões do dataset, nos modelos Random Forest, Decision Tree e XGBoost.

ajustados de forma equivalente, seguindo as recomendações das publicações originais, com o objetivo de garantir uma comparação justa e controlada entre as abordagens. A avaliação considerou a acurácia dos modelos após a remoção de diferentes proporções de dados, comparando os resultados obtidos com os alcançados por meio do retreinamento completo do modelo. Além disso, buscou-se compreender o impacto da taxa de remoção sobre o tempo e desempenho do retreinamento, com o intuito de identificar limitações e vantagens de cada abordagem em cenários de exclusão massiva de dados.

5. Resultados e Discussão

No **Experimento 1**, três modelos — Random Forest, Decision Tree e XGBoost — foram avaliados com o dataset MNIST, no qual cada instância foi associada a um identificador de usuário (*ownerID*) entre 0 e 9. Dois cenários foram considerados: IID, com distribuições uniformes das classes em proporções diferentes, e non-IID, com quantidades iguais de instâncias, mas distribuição não homogênea. A exclusão completa dos dados de um *ownerID* por vez simulou pedidos de remoção, permitindo analisar o impacto na acurácia dos modelos. Os resultados (Figura 2) indicaram que a quantidade de dados removidos não afetou de forma significativa o desempenho final, em ambos os cenários.

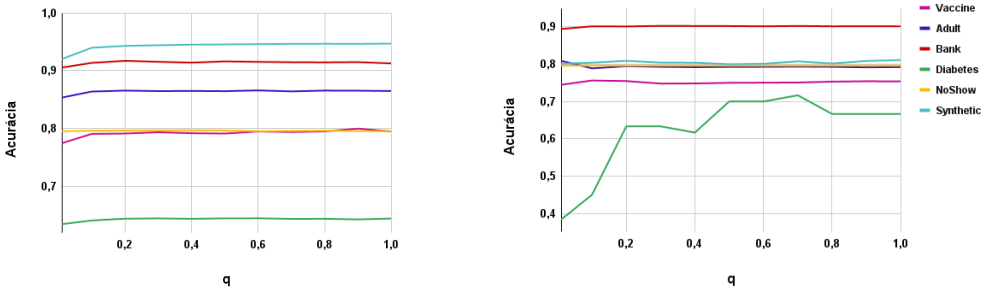
Tabela 2. Acurácia da remoção percentual no algoritmo DynFrs.

q	Vaccine	Adult	Bank	Diabetes	NoShow	Synthetic
0.01	**0.7652**	**0.8439**	**0.9057**	**0.6156**	**0.7958**	**0.9104**
0.10	0.7813	0.8543	0.9138	0.6319	0.7966	0.9299
0.20	0.7918	0.8660	0.9173	0.6449	0.7970	0.9431
0.30	0.7941	0.8651	0.9156	0.6455	0.7974	0.9442
0.40	0.7924	0.8654	0.9142	0.6447	0.7973	0.9453
0.50	0.7918	0.8650	0.9164	0.6455	0.7972	0.9456
0.60	0.7954	0.8663	0.9155	0.6456	0.7958	0.9461
0.70	0.7945	0.8646	0.9149	0.6446	0.7960	0.9466
0.80	0.7956	0.8660	0.9147	0.6447	0.7968	0.9467
0.90	0.8003	0.8658	0.9150	0.6437	0.7961	0.9466
1.00	0.7954	0.8652	0.9128	0.6452	0.7961	0.9470

Na sequência, para o **Experimento 2**, empregou-se o algoritmo DynFrs — fundamentado em florestas aleatórias — visando replicar os experimentos descritos no estudo

original de [Wang et al. 2025]. Os parâmetros configurados para o experimento foram: (i) T, representando o número de árvores na floresta, fixado em 100; (ii) k, definido como $k = \lceil qT \rceil$, de modo que cada amostra estivesse presente em no máximo k árvores. Como $q = k/T$, onde q corresponde à fração de amostras removidas por árvore, o parâmetro q variou entre 1%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% e 100% na remoção dos dados dos algoritmos Vaccine, Adult, Bank, Diabetes, NoShow e Synthetic, visando avaliar o comportamento do algoritmo.

Na Figura 3a, os resultados confirmaram as conclusões do estudo original: para a maioria dos conjuntos de dados, uma queda na acurácia foi observada apenas para $q < 0.1$, sendo mais acentuada no dataset Vaccine. Os resultados detalhados podem ser verificados na Tabela 2. O dataset NoShow não apresentou queda relevante, como pode ser visto em vermelho na tabela.



(a) DynFrs. **(b) DaRE.**

Figura 3. Acurácia da remoção percentual nos algoritmos DynFrs e DaRE

O algoritmo DaRE foi aplicado de forma comparável ao DynFrs, utilizando a mesma seleção aleatória de amostras para remoção, gerada pela biblioteca Numpy, garantindo reprodutibilidade experimental. A configuração empregou $n\_estimators$=100, max_depth=3, k=5 e $topd$=0, com a remoção de dados realizada pelo método nativo `drf.delete(indices)`, assegurando a execução de desaprendizado verdadeiro em vez de retreinamento completo. Os resultados mostraram que o DaRE apresentou comportamento semelhante ao DynFrs nos conjuntos NoShow, Synthetic, Vaccine e Bank, mas revelou deterioração significativa no dataset Diabetes. Já no Adult, o melhor desempenho ocorreu na remoção extrema (q=0.01), enquanto valores superiores resultaram em queda de acurácia, conforme evidenciado na Tabela 3 e Figura 3b.

Os experimentos demonstraram que o impacto da remoção de dados na acurácia dos modelos varia conforme o algoritmo e o conjunto de dados. Observou-se que o parâmetro q do método OCC(q) define um *trade-off* entre eficiência de desaprendizado e acurácia: valores baixos ($q < 0,1$) aceleram o processo por reduzir o número de subárvores afetadas, mas comprometem a diversidade do *ensemble* e a capacidade de generalização do modelo.

Por outro lado, valores mais altos ($q > 0,5$) preservam a diversidade do *ensemble* e mantêm alta acurácia, mas aumentam o custo computacional, já que cada remoção impacta um número maior de árvores. Assim, a configuração ótima de q depende do contexto: aplicações com solicitações frequentes de esquecimento favorecem valores baixos,

Tabela 3. Acurácia da remoção percentual no algoritmo DaRE.

q	Vaccine	Adult	Bank	Diabetes	NoShow	Synthetic
0.01	**0.7385**	**0.8091**	**0.8944**	**0.3833**	**0.7959**	**0.7913**
0.10	0.7463	0.7889	0.9016	0.4500	0.7959	0.7941
0.20	0.7448	0.7947	0.9012	0.6333	0.7959	0.8090
0.30	0.7480	0.7931	0.9026	0.6333	0.7959	0.8043
0.40	0.7484	0.7921	0.9024	0.6167	0.7959	0.8042
0.50	0.7501	0.7928	0.9022	0.7000	0.7959	0.8000
0.60	0.7505	0.7931	0.9016	0.7000	0.7959	0.8013
0.70	0.7508	0.7935	0.9024	0.7167	0.7959	0.8076
0.80	0.7533	0.7930	0.9016	0.6667	0.7959	0.8019
0.90	0.7542	0.7917	0.9017	0.6667	0.7959	0.8086
1.00	0.7538	0.7924	0.9016	0.6667	0.7959	0.8112

enquanto cenários que priorizam máxima acurácia devem adotar valores elevados, mesmo com maior latência no processo de desaprendizado.

6. Conclusão

Os resultados obtidos confirmam os objetivos propostos, evidenciando que a remoção de pequenas quantidades de dados exerce impacto limitado na acurácia de modelos tradicionais, funcionando como baseline para análises posteriores. A avaliação dos algoritmos DaRE e DynFrs demonstrou que a eficácia e a eficiência do desaprendizado variam conforme o algoritmo e o conjunto de dados, ressaltando a necessidade de estratégias adaptativas a diferentes cenários. Além disso, a capacidade de executar desaprendizado com degradação mínima de acurácia (¡1%) confirma a viabilidade técnica de garantir o direito ao esquecimento previsto em legislações como LGPD e GDPR, sem comprometer significativamente a utilidade prática dos modelos.

A partir desses achados, destacam-se duas direções promissoras para trabalhos futuros: a validação do processo de desaprendizado por meio de tecnologias blockchain e a integração com técnicas de inteligência artificial explicável (XAI). O uso de blockchain pode oferecer registros imutáveis e rastreáveis das operações de esquecimento, garantindo transparência e conformidade legal perante órgãos reguladores. Por sua vez, a incorporação de XAI favorece a interpretação e a confiabilidade das decisões algorítmicas após o desaprendizado, assegurando que os sistemas permaneçam transparentes e controláveis mesmo diante da exclusão de dados sensíveis.

Apesar dos avanços, algumas limitações devem ser consideradas. A eficácia dos algoritmos depende fortemente das características dos dados e do contexto de aplicação, o que indica a necessidade de explorar metodologias mais generalizáveis. Além disso, a sobrecarga computacional introduzida por mecanismos adicionais, como blockchain e XAI, requer investigações futuras sobre escalabilidade e eficiência. Esses pontos abrem espaço para pesquisas que conciliem desaprendizado seguro, explicabilidade e rastreabilidade em ambientes de larga escala, promovendo soluções de IA mais éticas, auditáveis e socialmente responsáveis.

Agradecimentos

Ao programa PIIC/Ufes, pela bolsa de estudos de IC durante toda a pesquisa. À Fapes, CNPq pelo financiamento parcial por meio de projetos de pesquisa do DSL/PPGI/Ufes. Ae ao GT-Audita/RNP, pelo apoio técnico no desenvolvimento deste trabalho.

Referências

Alpaydin, E. and Kaynak, C. (1998). Optical Recognition of Handwritten Digits. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50P49.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. (2021). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

Brophy, J. and Lowd, D. (2021). Machine unlearning for random forests. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1092–1104. PMLR.

Dang, Q.-V. (2021). Right to be forgotten in the age of machine learning. In Antipova, T., editor, *Advances in Digital Science*, pages 403–411, Cham. Springer International Publishing.

Jin, R., Chen, M., Zhang, Q., and Li, X. (2024). Forgettable federated linear learning with certified data unlearning.

Li, Y., Chen, C., Zhang, Y., Liu, W., Lyu, L., Zheng, X., Meng, D., and Wang, J. (2023). UltraRE: Enhancing receraser for recommendation unlearning via error decomposition. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nguyen, T. T., Huynh, T. T., Ren, Z., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. (2024). A survey of machine unlearning.

Presidência da República (2018). Lei geral de proteção de dados pessoais (lgpd). `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`. Acesso em: 24 jun. 2025.

Schelter, S., Grafberger, S., and Dunning, T. (2021). Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21)*, page –, Virtual Event, China.

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. (2022). Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319.

União Europeia (2016). Regulamento (ue) 2016/679 do parlamento europeu e do conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados (regulamento geral sobre a proteção de dados - rgpd). `https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=celex%3A32016R0679`. Acesso em: 21 jun. 2024.

Wang, S., Shen, Z., Qiao, X., Zhang, T., and Zhang, M. (2025). Dynfrs: An efficient framework for machine unlearning in random forest.