FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Aplicação e Avaliação de Aprendizagem Automática com Preservação da Privacidade em Sistemas Bancários

João Nunes Gonçalves

Mestrado em Inteligência Artificial

Orientador: Prof. Miriam Santos

Co-orientador: Prof. Ricardo Pereira

29 Julho, 2026

Aplicação e Avaliação de Aprendizagem Automática com Preservação da Privacidade em Sistemas Bancários

João Nunes Gonçalves

Mestrado em Inteligência Artificial

Aprovado em provas públicas pelo Júri:

Presidente: Inês Dutra
Arguente: Tânia Carvalho
Orientador: Miriam Seoane Santos

29 Julho, 2026

Resumo

A adoção de modelos de Aprendizagem Automática em sistemas de avaliação de risco de crédito bancário introduz riscos de privacidade que transcendem a proteção de identificadores diretos. Modelos treinados sobre dados financeiros sensíveis podem expor informação individual mesmo sem acesso aos dados originais, através de ataques que determinam se um indivíduo pertenceu ao conjunto de treino ou que recuperam atributos sensíveis a partir das previsões do modelo.

Esta dissertação avalia empiricamente um conjunto alargado de técnicas de aprendizagem automática com preservação de privacidade num processo real de avaliação de risco de crédito, desenvolvido em colaboração com a ITSCredit. São testadas cinco famílias de métodos (mecanismos de privacidade diferencial local, geração de dados sintéticos com privacidade diferencial, proteção do modelo durante o treino, perturbação simples e perturbação seletiva), avaliadas através de métricas de utilidade e de privacidade adequadas ao contexto.

Os resultados mostram que a fuga de informação por inferência de atributos é estrutural e proporcional ao poder preditivo das variáveis, impondo um limite teórico à redução de privacidade alcançável sem sacrificar a utilidade. A perturbação seletiva calibrada pela fuga de informação efetiva de cada variável revela-se a abordagem mais eficaz, com o *Generalized Randomized Response* Seletivo (GRR Seletivo) a alcançar uma redução de 25.1% na soma da fuga de informação nas quatro variáveis de maior fuga de informação na Regressão Logística e de 35.8% no *CatBoost*, com garantias formais de privacidade diferencial local.

Agradecimentos

À Professora Miriam Santos e ao Professor Ricardo Pereira, orientadores desta dissertação, pelo acompanhamento, disponibilidade e orientação ao longo de todo o trabalho. O profissionalismo e esforço de ambos, as discussões, sugestões e partilha de conhecimento foram essenciais para o desenvolvimento e melhoria desta dissertação.

Ao Micael Santos e ao Paulo Pinto, da ITSCredit, pela disponibilização dos dados, pelo acompanhamento próximo e por tornarem possível a realização deste trabalho num ambiente real.

João Nunes Gonçalves

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CAT	CatBoost
CTGAN	Conditional Tabular Generative Adversarial Network
DCR	Distance to Closest Record
DP	Differential Privacy — Privacidade Diferencial
DPGAN	Differentially Private Generative Adversarial Network
DP-SGD	Differentially Private Stochastic Gradient Descent
DP FT SD	Differential Privacy Frequency Tables Synthetic Data
GAN	Generative Adversarial Network
GRR	Generalized Randomized Response
KS	Kolmogorov-Smirnov
LDP	Local Differential Privacy — Privacidade Diferencial Local
LR	Logistic Regression — Regressão Logística
MIA	Membership Inference Attack
ML	Machine Learning — Aprendizagem Automática
NNDR	Nearest Neighbour Distance Ratio
OUE	Optimized Unary Encoding
PATE	Private Aggregation of Teacher Ensembles
PPML	Privacy-Preserving Machine Learning
RAPPOR	Randomized Aggregatable Privacy-Preserving Ordinal Response
RCS	Random Category Substitution
RF	Random Forest
RGPD	Regulamento Geral sobre a Protecção de Dados
WOE	Weight of Evidence — Peso de Evidência

Capítulo 1

Introdução

Este capítulo introduz o problema de investigação que motiva esta dissertação, contextualizando a crescente adoção de técnicas de *Machine Learning* na avaliação de risco de crédito e os riscos de privacidade que daí decorrem, definindo os objetivos geral e específicos da dissertação e descrevendo a estrutura do restante documento.

1.1 Contexto e Motivação

Nos últimos anos, o setor bancário tem vindo a incorporar de forma crescente técnicas de Inteligência Artificial (IA) e *Machine Learning (ML)* para apoiar decisões ao longo do ciclo de vida do crédito, desde a análise de risco até à gestão de incumprimentos [1–3]. Modelos de avaliação de crédito baseados em dados históricos permitem estimar a probabilidade de incumprimento de um cliente, contribuindo para decisões mais informadas sobre concessão de crédito, definição das condições financeiras (taxas, comissões) e políticas de risco.

Este avanço tecnológico depende, porém, do tratamento intensivo de dados bancários de clientes, que constituem uma das categorias de dados pessoais mais sensíveis [4]. A utilização destes dados é fortemente condicionada por enquadramentos legais e normativos, como o Regulamento Geral de Proteção de Dados (RGPD), o DORA (*Digital Operational Resilience Act*) e orientações de autoridades de supervisão e de proteção de dados [5]. Conciliar o potencial dos modelos de ML com a necessidade de garantir confidencialidade, minimização de dados e uma abordagem de proteção de dados desde a conceção (*privacy-by-design*) tornou-se um desafio central em sistemas de decisão automatizada em banca e um fator crítico para a confiança de clientes e reguladores.

A área de *Privacy-Preserving Machine Learning* (PPML) surge neste contexto como um conjunto de técnicas que visam permitir a utilização de dados em modelos de ML reduzindo o risco de exposição de informação sensível [6]. Estas técnicas podem atuar em diferentes níveis do sistema: (i) ao nível dos dados, através de anonimização, agregação, perturbação ou publicação diferencialmente privada; (ii) ao nível do modelo e do processo de treino, com mecanismos como Privacidade Diferencial, algoritmos robustos a ataques de inferência ou defesa contra *membership*

inference; (iii) ao nível arquitetural, com abordagens como *federated learning*, computação multi-partidária segura ou *Trusted Execution Environments* (TEE); e (iv) em configurações híbridas, que combinam elementos de mais do que uma destas famílias.

Apesar da evolução significativa da literatura nestas diferentes frentes, a aplicação prática de PPML a fluxos reais de classificação de crédito continua pouco explorada [6, 7]. Em particular, levantam-se questões sobre quais as famílias de técnicas mais adequadas a dados tabulares desbalanceados, processos de *binning* e seleção de um número reduzido de variáveis, requisitos de interpretabilidade e necessidade de integração com procedimentos e infraestruturas já estabelecidos no setor bancário.

1.2 Objetivos

Esta dissertação pretende analisar e comparar, de forma sistemática, diferentes famílias de técnicas de PPML num contexto real de avaliação de crédito, identificando quais as abordagens mais adequadas e quais os compromissos que surgem em termos de privacidade e desempenho.

Assim, este trabalho pretende estudar, em várias famílias de PPML, as técnicas mais relevantes para o contexto de avaliação do risco de crédito com dados tabulares, avaliar empiricamente o seu impacto quando aplicadas a um fluxo representativo deste tipo de sistemas e, com base nessa análise, propor uma arquitetura que agregue as abordagens mais adequadas para proteção de dados em modelos de avaliação de crédito num contexto real de uma empresa do sector bancário.

Este objetivo geral desdobra-se nos seguintes objetivos específicos:

1. Avaliar empiricamente um conjunto de técnicas representativas de *Privacy-Preserving Machine Learning* (PPML), pertencentes a diferentes famílias, em cenários de classificação de crédito com dados tabulares, medindo o impacto em métricas de desempenho como a AUC ROC (*Area Under the Receiver Operating Characteristic Curve*), amplamente utilizada para avaliar a capacidade discriminativa global de classificadores binários, e a estatística KS (*Kolmogorov–Smirnov*), que quantifica a distância máxima entre as distribuições acumuladas das classes de incumprimento e não incumprimento, sendo uma métrica tradicional no setor financeiro.

2. Quantificar os compromissos (*trade-offs*) entre privacidade e utilidade dos modelos dessas técnicas, incluindo a análise de diferentes configurações de proteção de privacidade.

3. Identificar, com base nos resultados experimentais, as abordagens de PPML ou combinações de abordagens que oferecem compromissos mais favoráveis para avaliação do risco de crédito, justificando a sua adequação ao contexto estudado.

4. Conceber, implementar e validar uma arquitetura integrada de PPML para avaliação do risco de crédito, que incorpore as abordagens mais promissoras identificadas e que possa servir como referência para a adoção prática de mecanismos de preservação de privacidade em sistemas de decisão de crédito.

O âmbito da dissertação centra-se em dados tabulares de risco de crédito e em fluxos de avaliação do risco de crédito que reflitam práticas comuns no setor bancário, servindo como cenário de teste para a comparação entre famílias de PPML. Esta investigação enquadra-se no contexto de uma dissertação empresarial desenvolvida em colaboração com a ITSCredit, recorrendo a dados reais disponibilizados pela organização e a requisitos e práticas alinhadas com o seu contexto operacional.

1.3 Estrutura do Documento

O presente documento está organizado em cinco capítulos, cuja estrutura reflete a progressão metodológica seguida ao longo do trabalho:

- O Capítulo 1 contextualiza o problema, apresenta a motivação para a investigação e define os objetivos específicos a atingir.
- O Capítulo 2 apresenta a revisão da literatura, cobrindo os fundamentos de classificação de crédito, os riscos de privacidade em *Machine Learning* e as principais famílias de técnicas de PPML documentadas na literatura, culminando numa síntese dos trabalhos relacionados mais relevantes para este contexto.
- O Capítulo 3 descreve os dados utilizados, o fluxo de processamento adotado e os modelos de base, bem como as métricas de avaliação de utilidade e de privacidade que orientam toda a experimentação.
- O Capítulo 4 apresenta a análise experimental, organizada por famílias de métodos: mecanismos de Privacidade Diferencial Local, geração de dados sintéticos com Privacidade Diferencial, proteção do modelo durante o treino, perturbação por ruído simples e perturbação seletiva. Para cada método são apresentados e discutidos os resultados obtidos em termos de utilidade e privacidade.
- O Capítulo 5 sintetiza os resultados, apresenta uma análise comparativa entre famílias de métodos, discute o compromisso privacidade-utilidade, propõe uma arquitetura integrada para adoção pela ITSCredit e identifica as limitações do trabalho e as direções de investigação futura.

Capítulo 2

Revisão da Literatura

Este capítulo apresenta o enquadramento teórico que sustenta a investigação desenvolvida nesta dissertação, cobrindo a classificação de crédito como problema de *Machine Learning* e os riscos de privacidade associados, as principais famílias de técnicas de *Privacy-Preserving Machine Learning* documentadas na literatura e o seu posicionamento face aos trabalhos mais próximos do contexto desta dissertação.

2.1 Classificação de crédito: abordagens, privacidade e riscos

A classificação de crédito é frequentemente concetualizada como um problema de classificação binária, no qual se estima a probabilidade de incumprimento (*default*) a partir de variáveis demográficas, comportamentais e financeiras do cliente [8]. Historicamente, modelos estatísticos lineares, em particular a regressão logística, foram amplamente utilizados na avaliação do risco de crédito, em parte por oferecerem relações mais diretamente interpretáveis entre variáveis e risco, o que facilita processos de validação, auditoria e supervisão [9, 10]. A construção de tabelas de pontuação de crédito baseados em regressão logística segue tipicamente um processo estruturado que inclui a preparação dos dados, discretização (*binning*) de variáveis contínuas e codificação das variáveis categóricas com métricas como *Weight of Evidence* (WoE) e *Information Value* (IV) [8, 11]. Estas práticas são amplamente documentadas na literatura aplicada de classificação de crédito e em guias técnicos, sobretudo por contribuírem para modelos mais estáveis e interpretáveis no contexto de decisão de crédito [11].

Nas últimas duas décadas, observa-se uma adoção crescente de métodos de *Machine Learning* em classificação de crédito, acompanhando a maior disponibilidade de dados e recursos computacionais, bem como o interesse em capturar não linearidades e interações entre variáveis [12, 13]. Revisões sistemáticas recentes documentam um aumento significativo na utilização de árvores de decisão, *random forests*, *gradient boosting machines* (incluindo variantes como XGBoost, LightGBM e CatBoost), máquinas de vetores de suporte e redes neuronais na modelação do risco de crédito [9, 12, 13]. Outros estudos comparativos indicam que modelos baseados em *ensembles*

de árvores (por exemplo, *gradient boosting*) frequentemente superam modelos lineares em métricas como AUC e coeficiente de Gini, sobretudo quando existem relações não lineares e interações complexas entre variáveis [12, 13].

Apesar destes avanços, a adoção de modelos de maior complexidade em instituições financeiras é frequentemente condicionada por exigências de governação, validação e controlo de alterações do modelo, bem como por expectativas de supervisão relativamente à robustez e monitorização do desempenho [10, 14]. Diversos autores salientam que, mesmo quando se recorre a algoritmos mais sofisticados, é frequente manter práticas inspiradas nas tabelas de pontuação de crédito tradicionais (como o recurso a discretização supervisionada, análise de *Weight of Evidence* e validações segmentadas) de forma a preservar a interpretabilidade e facilitar a validação por equipas de risco e auditoria [12, 13]. Este equilíbrio entre desempenho e explicabilidade é particularmente relevante quando se pretende introduzir mecanismos adicionais de preservação de privacidade num sistema de avaliação do risco de crédito, tema que será explorado nas secções seguintes.

2.1.1 Privacidade e proteção de dados em contexto bancário

A atividade bancária assenta, de forma estrutural, no tratamento de grandes volumes de dados pessoais, incluindo categorias particularmente sensíveis do ponto de vista económico e social, como históricos de crédito, informação sobre rendimentos, movimentos de conta ou situações de incumprimento [1]. Neste contexto, a proteção de dados deixa de ser apenas uma obrigação legal abstrata e torna-se um fator central de confiança entre clientes e instituições financeiras, bem como um elemento crítico para a estabilidade e reputação do sistema financeiro.

No espaço europeu, o enquadramento normativo é dominado pelo Regulamento Geral sobre a Proteção de Dados (RGPD), que estabelece um conjunto de princípios basilares para qualquer tratamento de dados pessoais: licitude, lealdade e transparência; limitação das finalidades; minimização dos dados; exatidão; limitação da conservação; integridade e confidencialidade; e responsabilização [4, 15]. Estes princípios têm implicações diretas para casos de uso como o classificação de crédito: a instituição deve justificar a base legal para o tratamento (por exemplo, cumprimento de obrigação legal ou interesse legítimo), limitar os dados aos estritamente necessários para avaliar a solvabilidade, garantir a segurança técnica e organizativa dos sistemas e ser capaz de demonstrar conformidade perante a autoridade de supervisão.

Para além do RGPD, o setor bancário está sujeito a regulamentação específica, que articula prudência financeira com proteção de dados. As *Guidelines on Loan Origination and Monitoring* da Autoridade Bancária Europeia (EBA) sublinham a necessidade de critérios robustos de avaliação de solvência e de governação interna, incluindo requisitos sobre qualidade, relevância e proporcionalidade dos dados utilizados na concessão de crédito [16]. Em paralelo, o Supervisor Europeu de Proteção de Dados (EDPS) tem chamado a atenção para os riscos associados ao uso extensivo de dados pessoais e de perfis automatizados em crédito ao consumo, recomendando a clarificação das categorias de dados admissíveis, a limitação do recurso a fontes externas e maior transparência sobre decisões automatizadas [17].

As autoridades nacionais de proteção de dados, como a Comissão Nacional de Proteção de Dados (CNPD) em Portugal, têm igualmente salientado a importância de avaliar cuidadosamente os riscos de privacidade associados a sistemas de IA e de avaliação de risco, incluindo questões de opacidade algorítmica, potenciais efeitos discriminatórios e a necessidade de supervisão humana adequada [18, 19]. Em paralelo, o recente Regulamento Europeu em matéria de Inteligência Artificial (AI Act) classifica sistemas de IA utilizados para avaliação de crédito como sistemas de alto risco, reforçando exigências de gestão de risco, qualidade de dados, transparência e governança [20].

Neste ambiente regulatório, o desenvolvimento de modelos de *Machine Learning* para classificação de crédito não pode ser visto apenas como um problema técnico de otimização de métricas preditivas. As instituições devem demonstrar que os dados tratados são adequados, pertinentes e limitados ao necessário, que existem medidas técnicas e organizativas para mitigar riscos de divulgação indevida e potenciais cenários de reidentificação, e que os modelos não agravam desigualdades ou práticas discriminatórias [4, 20, 21]. É precisamente neste cruzamento entre inovação analítica e exigências regulatórias que surge a necessidade de recorrer a técnicas de *Privacy-Preserving Machine Learning*, procurando conciliar utilidade dos modelos com garantias robustas de privacidade e conformidade normativa.

2.1.2 Riscos de privacidade em *Machine Learning*

A utilização de modelos de *Machine Learning* em contexto bancário levanta riscos específicos de privacidade que vão além das preocupações clássicas de segurança da informação (por exemplo, controlo de acessos, gestão de credenciais ou encriptação). Mesmo quando os dados são tratados num ambiente controlado, o próprio modelo treinado pode, em determinadas condições, revelar informação sensível sobre os exemplos usados no treino [22]. A literatura documenta várias classes de ataques que exploram este fenómeno, incluindo ataques de inferência de pertença, de inversão de modelo, de inferência de atributos e de extração de dados de treino [23].

Os ataques de *membership inference* procuram determinar se um determinado registo foi, ou não, utilizado no conjunto de treino de um modelo. Shokri et al. [24] mostram que, para modelos de classificação treinados com dados sensíveis, é possível construir ataques que, a partir de acesso caixa-negra às probabilidades de saída do modelo, isto é, sem conhecimento interno da sua estrutura ou parâmetros, conseguem inferir a pertença de exemplos individuais ao conjunto de treino. Este tipo de ataque é particularmente relevante em cenários de avaliação de risco de crédito, nos quais a pertença ao conjunto de treino pode revelar, indiretamente, a existência de uma relação bancária ativa ou a exposição a eventos de risco (por exemplo, episódios de incumprimento).

Os ataques de inversão de modelo procuram, a partir do acesso ao modelo treinado (e eventualmente às suas saídas probabilísticas), reconstruir informação aproximada sobre os dados usados no treino ou inferir atributos privados dos indivíduos. Fredrikson, Jha e Ristenpart [25] demonstram que, em modelos de regressão e redes neuronais treinados com dados sensíveis, é possível inferir atributos privados de indivíduos a partir de outros atributos observáveis e das respostas do modelo. Em contextos de crédito, isto sugere que um modelo disponibilizado via API pode, em

princípio, ser explorado para inferir informação financeira não diretamente observável, caso não existam mecanismos de mitigação e restrições de acesso/consulta.

Relacionado com os ataques de inversão, mas conceptualmente distinto, o *Attribute Inference* constitui um risco específico em contextos onde os dados combinam variáveis preditivas e demograficamente sensíveis. Neste tipo de ataque, o adversário conhece todos os atributos de um indivíduo com exceção de um, e utiliza as saídas probabilísticas do modelo para inferir o valor desse atributo sensível [26]. Yeom et al. [26] mostram formalmente que a capacidade de inferência está diretamente relacionada com a influência do atributo nas previsões do modelo, o que implica que as variáveis com maior poder preditivo são simultaneamente as mais vulneráveis a este tipo de ataque.

Para além destes ataques direcionados, estudos sobre memorizações não intencionais em modelos de grande capacidade mostram que redes neuronais podem reter e expor segmentos de dados de treino quase literalmente, sobretudo quando expostas através de interfaces que permitem consultas repetidas a distribuições de probabilidade ou processos de geração [23, 27]. Embora muitos destes trabalhos incidam sobre modelos de linguagem, o princípio geral, de que modelos sobreparametrizados podem memorizar exemplos raros ou atípicos, é relevante também para sistemas de pontuação de crédito treinados com dados bancários potencialmente altamente identificáveis.

A estes riscos associados ao modelo juntam-se riscos mais tradicionais de reidentificação em conjuntos de dados previamente anonimizados. Trabalhos clássicos em anonimização de dados mostram que a combinação de poucos atributos quase-identificadores (como idade, código postal e género) pode ser suficiente para reidentificar indivíduos em bases de dados supostamente anónimas [28, 29]. Em aplicações de avaliação de risco de crédito, onde coexistem variáveis demográficas, financeiras e comportamentais, o risco de reidentificação tende a aumentar, sobretudo quando conjuntos de dados são partilhados com terceiros para desenvolvimento, auditoria ou validação de modelos.

Em síntese, a literatura evidencia que: (i) modelos de *Machine Learning* podem tornar-se vetores de fuga de informação sobre os dados de treino; (ii) ataques de inferência de pertença, de atributos e de inversão podem ser viáveis sob pressupostos práticos (por exemplo, acesso caixa-negra ao modelo); e (iii) técnicas tradicionais de anonimização baseadas em generalização e supressão podem ser insuficientes. Estes resultados motivam o recurso a abordagens de *Privacy-Preserving Machine Learning* que ofereçam garantias formais ou, pelo menos, mecanismos sistemáticos de redução do risco de divulgação em sistemas de avaliação de risco de crédito.

2.2 *Privacy-Preserving Machine Learning*

O termo *Privacy-Preserving Machine Learning* (PPML) designa um conjunto de técnicas cujo objetivo é permitir o treino e a utilização de modelos de *Machine Learning* com dados sensíveis, reduzindo o risco de reidentificação de indivíduos ou de inferência de informação privada. Em contraste com abordagens tradicionais baseadas apenas em anonimização ou controlo de acesso, a PPML assume explicitamente que os próprios modelos, os parâmetros treinados ou estatísticas

intermédias podem expor informação privada, por via de ataques como *membership inference*, *model inversion* e reconstrução de dados de treino [30].

A literatura organiza este conjunto de técnicas segundo taxonomias que refletem (i) em que ponto do sistema a privacidade é garantida e (ii) que pressupostos de confiança são introduzidos (por exemplo, confiança numa entidade central de agregação, em *hardware* seguro, ou ausência de confiança em qualquer entidade central) [30, 31]. Neste trabalho, adota-se uma caracterização por camadas, distinguindo-se técnicas (i) ao nível dos dados, (ii) ao nível do modelo e do processo de treino e (iii) ao nível arquitetural.

Ao nível dos dados, incluem-se técnicas como generalização e supressão (por exemplo, *k-anonymity* e extensões), amplamente utilizadas para reduzir o risco de reidentificação em dados tabulares [32]; mecanismos de perturbação aleatória, que introduzem ruído controlado nos atributos para limitar a informação individual preservando, tanto quanto possível, propriedades estatísticas úteis [33]; *Local Differential Privacy*, em que a perturbação é aplicada do lado do utilizador antes da recolha centralizada dos dados [34]; e geração de dados sintéticos com objetivos explícitos de privacidade, procurando disponibilizar dados substitutos com utilidade analítica e menor risco de divulgação [35].

Ao nível do modelo e do processo de treino, destacam-se abordagens como *Differential Privacy*, aplicada a gradientes ou atualizações de parâmetros [36], que introduz ruído controlado no processo de otimização para limitar a contribuição de cada registo individual. Complementarmente, surgem métodos de treino distribuído que reduzem a exposição de dados e parâmetros durante a aprendizagem, como o *privacy-preserving deep learning* proposto por Shokri e Shmatikov [37], no qual múltiplas entidades treinam modelos locais e partilham apenas subconjuntos selecionados de atualizações de parâmetros com um servidor agregador, mitigando a necessidade de centralizar dados em claro.

Ao nível arquitetural, enquadram-se técnicas que reconfiguram a forma como os dados e o modelo são distribuídos entre as partes envolvidas, como aprendizagem federada com agregação segura, *Secure Multiparty Computation* (SMC) e esquemas baseados em encriptação homomórfica ou *Trusted Execution Environments* (TEE) [31, 38]. Estas abordagens visam minimizar a necessidade de centralizar dados sensíveis e, em muitos casos, podem ser combinadas com *Differential Privacy*, dando origem a soluções híbridas que procuram equilibrar privacidade, desempenho do modelo e custo computacional [39]. Neste trabalho, a experimentação incide sobre técnicas ao nível dos dados e ao nível do modelo e do processo de treino, sendo a camada arquitetural abordada apenas no contexto da revisão da literatura e dos trabalhos relacionados.

2.2.1 PPML ao nível dos dados

As técnicas de *Privacy-Preserving Machine Learning* ao nível dos dados procuram reduzir o risco de divulgação de informação sensível antes mesmo do treino do modelo, transformando o conjunto de dados de forma a que este possa ser utilizado com menor risco de reidentificação ou de inferência de atributos. Em termos gerais, estas abordagens atuam sobre os registos ou sobre estatísticas agregadas. Neste trabalho, seguindo uma organização inspirada em revisões da

literatura sobre técnicas de proteção de dados em aprendizagem automática [40, 41], distinguem-se quatro grandes famílias: (i) anonimização clássica por generalização e supressão, amplamente documentada na literatura como ponto de partida histórico da área; (ii) perturbação aleatória com garantias formais de *Local Differential Privacy*; (iii) perturbação heurística sem garantias formais de privacidade; e (iv) publicação e geração de dados sintéticos com incorporação de mecanismos de *Differential Privacy* central. As famílias (ii), (iii) e (iv) são as avaliadas experimentalmente neste trabalho.

A anonimização clássica assenta na ideia de que os registos individuais não devem ser distinguíveis dentro de um grupo suficientemente grande de indivíduos com atributos quase-identificadores idênticos. O modelo de *k-anonymity*, introduzido por Sweeney [28], formaliza esta noção exigindo que cada combinação de atributos quase-identificadores apareça em pelo menos k registos do conjunto de dados. Para atingir esta propriedade recorre-se tipicamente a generalização (por exemplo, transformar idades exatas em classes etárias) e supressão de valores raros. No entanto, trabalhos posteriores mostraram que o *k-anonymity* é vulnerável em contextos em que a distribuição dos atributos sensíveis dentro de um grupo é muito homogénea: um atacante que conheça os quase-identificadores de um indivíduo pode inferir, com elevada confiança, o seu atributo sensível. Para mitigar esta limitação foram propostos modelos como l-diversity [42] e t-closeness [43], que exigem diversidade de valores sensíveis e proximidade à distribuição global, respetivamente. Embora amplamente estudadas, estas técnicas enfrentam dificuldades práticas em dados tabulares de elevada dimensionalidade: à medida que aumenta o número de quase-identificadores, torna-se mais provável que combinações de atributos sejam raras, obrigando a generalizar ou suprimir uma fração significativa dos valores para satisfazer k. Esse processo reduz a granularidade dos atributos e pode distorcer relações estatísticas relevantes, levando a perdas de utilidade para tarefas de *Machine Learning* [44].

A segunda família baseia-se na perturbação aleatória com garantias formais de privacidade, enquadrada no paradigma de *Local Differential Privacy* [45]. Neste modelo, cada registo é transformado de forma que a contribuição individual se torne indistinguível dentro de um certo nível de ruído, com garantia formal calibrada pelo parâmetro ε. O mecanismo de *randomized response*, introduzido originalmente em contexto de ensaios aleatorizados [46], fornece a ideia base: em vez de reportar diretamente o valor verdadeiro, o respondente aplica uma transformação aleatória controlada, que garante plausível negabilidade mas permite ao analista recuperar estatísticas agregadas quase não enviesadas. Extensões como o *Generalized Randomized Response* (GRR) e a *Optimized Unary Encoding* (OUE) adaptam estes princípios a variáveis categóricas e a codificações binárias de alta dimensionalidade [34, 45, 47]. O *Randomized Aggregatable Privacy-Preserving Ordinal Response* (RAPPOR), desenvolvido pela Google para recolha de estatísticas de utilização em larga escala, introduz uma dupla camada de perturbação sequencial que visa dificultar a correlação entre múltiplos reportes do mesmo indivíduo ao longo do tempo [34]. Em contexto de PPML, estas técnicas permitem construir conjuntos de dados já privatizados que podem ser usados em algoritmos de treino convencionais, à custa de alguma degradação do sinal estatístico.

A terceira família engloba abordagens heurísticas de perturbação categórica, sem calibração

teórica formal mas computacionalmente simples e de fácil aplicação em contextos operacionais. A substituição aleatória de categorias consiste em substituir o valor original de uma variável por uma categoria diferente escolhida aleatoriamente, com uma probabilidade p controlada pelo utilizador [48]. A troca de categorias entre registos é uma técnica amplamente utilizada em controlo de divulgação estatística, em que os valores de uma variável são permutados entre pares de registos, preservando a distribuição marginal de cada variável mas quebrando as associações entre registos [49]. Para variáveis numéricas, a perturbação por ruído gaussiano ou de Laplace consiste em adicionar a cada valor uma perturbação amostrada de uma distribuição centrada em zero, sendo uma abordagem clássica de mascaramento de dados contínuos [33]. Ao contrário dos mecanismos de *Local Differential Privacy*, estas abordagens não oferecem garantias quantificáveis sobre o nível de privacidade alcançado, mas são frequentemente utilizadas quando os requisitos regulatórios não exigem garantias formais ou quando a simplicidade de implementação é prioritária.

Finalmente, a quarta família incide sobre publicação de dados ou estatísticas com *Differential Privacy* central, geralmente sob a forma de tabelas de frequências, histogramas ou dados sintéticos [50]. No modelo central, assume-se a existência de um curador de confiança que observa os dados verdadeiros e aplica um mecanismo conforme à definição de *Differential Privacy*, introduzindo ruído calibrado em função de um parâmetro ε [51]. Com base em histogramas privatizados é possível divulgar estatísticas agregadas ou gerar amostras sintéticas que preservam, aproximativamente, as distribuições marginais e algumas dependências entre variáveis [52, 53]. Em alternativa, técnicas baseadas em modelos generativos podem ser treinadas com variantes de *Differential Privacy* no processo de otimização, produzindo dados sintéticos com garantias formais sobre a informação que um atacante pode inferir sobre qualquer indivíduo. O CTGAN (*Conditional Tabular GAN*) [54] é um modelo generativo adversarial desenvolvido especificamente para dados tabulares com variáveis mistas e categorias desbalanceadas, introduzindo um mecanismo de amostragem condicional que controla a distribuição das categorias geradas. A DPGAN (*Differentially Private GAN*) [55] incorpora *Differential Privacy* no processo de treino do discriminador através do mecanismo *Differentially Private Stochastic Gradient Descent* (DP-SGD), produzindo dados sintéticos com garantias formais de privacidade.

As abordagens ao nível dos dados apresentam várias vantagens em cenários industriais: são, em muitos casos, independentes do modelo de *Machine Learning* utilizado, podem ser aplicadas antes da partilha ou do treino do modelo e alinham-se naturalmente com princípios regulatórios como minimização de dados e proteção de dados desde a conceção. Contudo, a literatura evidencia também limitações importantes [6, 56, 57], em particular: (i) a dificuldade de assegurar simultaneamente forte privacidade e elevada utilidade em conjuntos de dados complexos e desbalanceados; (ii) a sensibilidade a ataques que combinem informação externa com o conjunto de dados transformado e/ou com estatísticas divulgadas; e (iii) a necessidade de analisar cuidadosamente a composição de privacidade quando múltiplas transformações ou mecanismos de proteção são aplicados sobre os mesmos dados. Embora cada mecanismo individual possa oferecer garantias formais de privacidade, a aplicação sucessiva de várias transformações pode levar à acumulação gradual de fuga de informação, reduzindo a proteção global assegurada ao conjunto de dados. Es-

tas questões são especialmente relevantes em sistemas de decisão de crédito, onde os dados são de elevada dimensionalidade, combinam atributos demográficos e financeiros e são utilizados para decisões com impacto significativo na vida dos clientes.

2.2.2 PPML ao nível do modelo e do treino

As abordagens de PPML ao nível do modelo e do processo de treino partem do princípio de que a própria dinâmica de aprendizagem pode ser controlada para limitar a influência de exemplos individuais no modelo final. Em vez de transformar os dados de entrada (como nas técnicas ao nível dos dados), estas abordagens modificam o algoritmo de treino, o objetivo de otimização e/ou o modo como atualizações e parâmetros são produzidos e expostos, reduzindo a informação observável por um adversário durante e após o treino.

Uma das linhas mais influentes nesta área assenta em *Differential Privacy* (DP) aplicada ao treino. Em particular, o algoritmo *DP-SGD* de Abadi et al. [36] adapta o gradiente estocástico padrão através de três componentes centrais: (i) truncagem do gradiente por exemplo, que limita a sensibilidade de cada passo de atualização a um único registo; (ii) adição de ruído gaussiano às somas de gradientes, calibrado ao nível de privacidade pretendido; e (iii) contabilização do orçamento de privacidade ao longo das iterações. Este mecanismo fornece garantias formais de DP (no modelo central), embora tipicamente implique degradação de desempenho que depende do tamanho do conjunto de dados, do número de iterações, da capacidade do modelo e do nível de privacidade imposto.

Em paralelo, diversos trabalhos exploram treino distribuído combinado com mecanismos de privacidade, aproximando-se da fronteira entre PPML ao nível do treino e PPML arquitetural. Shokri e Shmatikov [37] propõem um esquema de aprendizagem profunda com preservação de privacidade em que múltiplas entidades treinam localmente e partilham apenas partes das atualizações de gradientes com um servidor de agregação, mitigando a necessidade de centralizar dados brutos e reduzindo a informação diretamente exposta durante o treino. Embora esta proposta não forneça garantias formais de DP, evidencia como restringir a granularidade e a quantidade de informação partilhada pode reduzir riscos de inferência. Evoluções posteriores em aprendizagem federada introduzem mecanismos mais robustos, combinando agregação segura (para que o servidor não observe contribuições individuais) com DP para limitar, ao nível do cliente, a informação revelada pelas atualizações agregadas [38, 58].

Para além de DP, a literatura inclui ainda estratégias orientadas para reduzir a exposição do modelo e controlar o que é disponibilizado externamente. Uma classe relevante recorre a transferência de conhecimento para treinar um modelo aluno com base em informação agregada derivada de dados privados, evitando expor diretamente um modelo treinado sobre o conjunto sensível. A abordagem PATE (*Private Aggregation of Teacher Ensembles*) constrói um aluno a partir das previsões de múltiplos modelos professores treinados em partições disjuntas dos dados de treino, adicionando ruído ao processo de agregação de forma a satisfazer DP [59]. Estas abordagens são especialmente úteis quando é necessário disponibilizar uma interface de previsão pública, reduzindo o risco de que exemplos individuais sejam inferidos a partir das saídas do modelo.

Uma linha distinta de investigação aborda a equidade algorítmica como mecanismo de controlo sobre as previsões do modelo relativamente a variáveis sensíveis. Agarwal et al. [60] propõem uma abordagem de redução que transforma o problema de classificação com restrições de equidade numa sequência de problemas de classificação com custos, permitindo impor restrições como *DemographicParity* (independência estatística entre previsões e atributo sensível) e *EqualizedOdds* (igualdade das taxas de verdadeiros e falsos positivos entre grupos) de forma flexível e com garantias teóricas. O Fairlearn [61] é uma biblioteca que implementa esta abordagem, permitindo treinar classificadores com restrições explícitas de equidade sobre variáveis sensíveis definidas pelo utilizador. Embora a equidade algorítmica e a privacidade sejam objetivos conceptualmente distintos na literatura, a possibilidade de que restrições de equidade influenciem a fuga de informação associada às variáveis sensíveis justifica a sua inclusão neste contexto [62].

Em termos práticos, as técnicas ao nível do modelo e do treino são atrativas em cenários em que é desejável preservar a estrutura original dos dados e integrar a privacidade diretamente na governação do ciclo de vida do modelo, como em sistemas de avaliação de risco de crédito sujeitos a auditoria. Ainda assim, a literatura evidencia desafios: (i) degradação de desempenho para níveis de privacidade fortes, frequentemente mais acentuada em dados tabulares desbalanceados [63]; (ii) complexidade de implementação e calibração (por exemplo, limiar de truncagem, escala de ruído e contabilização de privacidade); e (iii) tensão entre estabilidade e explicabilidade e o ruído introduzido no treino. Estes compromissos são particularmente relevantes no setor bancário, onde o modelo deve ser simultaneamente eficaz, interpretável e conforme com requisitos de governação e proteção de dados.

2.2.3 PPML ao nível arquitetural

As abordagens de *Privacy-Preserving Machine Learning* (PPML) ao nível arquitetural centram-se na forma como os dados e os modelos são distribuídos, comunicados e processados entre diferentes entidades. Em vez de assumir a centralização dos dados de treino, estas técnicas reconfiguram a arquitetura do sistema para reduzir a necessidade de exposição de informação sensível e de confiança numa entidade única, recorrendo frequentemente a mecanismos criptográficos e/ou a *hardware* seguro.

Uma das linhas de trabalho mais marcantes neste domínio é a aprendizagem federada. Nesta configuração, os dados permanecem localmente em diferentes dispositivos ou instituições (por exemplo, diferentes bancos ou sucursais), e apenas atualizações do modelo, como gradientes ou parâmetros, são enviadas para um servidor de agregação que constrói um modelo global [7, 64]. O trabalho seminal de McMahan et al. [64] introduziu o algoritmo *Federated Averaging*, demonstrando que é possível treinar modelos em dados distribuídos, sem partilhar dados brutos. Contudo, a literatura reconhece que atualizações de modelo podem, em certos cenários, revelar informação sobre dados locais, motivando o uso de mecanismos adicionais. Desenvolvimentos posteriores incorporaram agregação segura [38] e, frequentemente, *Differential Privacy*, de forma a impedir que o servidor retenha contribuições individuais e a limitar a informação inferível a partir das atualizações agregadas.

Outra família de abordagens arquiteturais assenta em técnicas de computação segura, como a *Secure Multiparty Computation* (SMC) e a encriptação homomórfica. Em SMC, várias partes cooperam para calcular uma função conjunta sobre os seus dados privados, sem revelar os dados em claro umas às outras, sob diferentes modelos adversariais [65]. No modelo honesto mas curioso, assume-se que as entidades seguem corretamente o protocolo, mas tentam inferir informação adicional a partir das mensagens e do estado interno observado durante a execução [66]. Já no modelo malicioso, o adversário pode desviar-se do protocolo (por exemplo, forjar mensagens, omitir passos ou manipular entradas) para tentar extrair informação ou corromper o resultado, exigindo protocolos com mecanismos adicionais de verificação e robustez [65, 66]. A encriptação homomórfica, por sua vez, permite efetuar operações sobre dados encriptados, de forma que o resultado, após descodificação, seja equivalente ao cálculo que teria sido realizado sobre os dados originais [67]. Estas técnicas têm sido exploradas para treino e inferência em cenários de avaliação de risco de crédito, mas enfrentam desafios relevantes de eficiência, latência e complexidade de engenharia, sobretudo com modelos mais expressivos [31, 68].

Uma terceira abordagem arquitetural recorre a ambientes de execução confiável (*Trusted Execution Environment*, TEE), que disponibilizam enclaves seguros, isto é, regiões isoladas de execução protegidas ao nível do *hardware*, nas quais código e dados podem ser processados com garantias de confidencialidade e integridade. Nestes sistemas, o código de treino ou inferência é executado numa zona isolada do processador, com garantias de integridade e confidencialidade face ao sistema operativo e a outros processos [69]. Os dados sensíveis podem ser descodificados apenas dentro do enclave, processados pelo modelo e descartados, reduzindo o risco de exposição em memória ou em disco. Embora estas soluções não ofereçam garantias formais do tipo *Differential Privacy*, são apelativas em contextos regulados quando é possível controlar a infraestrutura e manter compatibilidade com modelos existentes.

As abordagens arquiteturais são particularmente relevantes em cenários multi-institucionais, em que diferentes entidades desejam cooperar no treino de modelos de risco sem partilhar diretamente os seus dados, ou quando a centralização é legalmente limitada. No entanto, a literatura destaca desafios importantes: (i) integração em fluxos de processamento industriais; (ii) dependência de modelos de ameaça e pressupostos (por exemplo, servidor honesto mas curioso ou confiança no *hardware*); e (iii) custo computacional e operacional associado a soluções criptográficas e/ou infraestrutura especializada [7]. Em avaliação de risco de crédito, estas abordagens tendem a ser complementares às técnicas ao nível dos dados e do treino, viabilizando, por exemplo, treino colaborativo entre instituições e execução mais segura em produção.

2.2.4 Métricas e compromissos de avaliação em PPML

A avaliação de técnicas de *Privacy-Preserving Machine Learning* (PPML) envolve, de forma inevitável, a análise de compromissos entre pelo menos três dimensões: (i) privacidade, (ii) utilidade dos modelos e (iii) custo computacional e operacional. A privacidade pode ser quantificada através de garantias formais (quando disponíveis), por exemplo *Differential Privacy* (DP), ou através de métricas empíricas que estimam o risco de divulgação (por exemplo, sucesso de ataques de

inferência). A utilidade é tipicamente medida por métricas de desempenho preditivo e/ou qualidade estatística (por exemplo, discriminação e calibração), enquanto o custo reflete o impacto em tempo de treino e inferência, consumo de recursos (CPU/GPU, memória, comunicação) e complexidade de integração em fluxos de processamento reais [56].

No caso de mecanismos que fornecem garantias formais, como em *Differential Privacy*, o nível de privacidade é controlado por parâmetros de um orçamento de privacidade, frequentemente descritos por (ε, δ), em que ε quantifica o nível máximo de fuga de informação permitido e δ representa a probabilidade de essa garantia falhar, sendo tipicamente um valor muito pequeno e próximo de zero [56]. De forma intuitiva, valores mais baixos de ε traduzem, em geral, uma proteção mais forte, mas tendem a exigir a injeção de mais ruído, o que pode degradar a utilidade. Na prática, a escolha de ε e de δ depende do mecanismo e do contexto: sensibilidade da função, modelo de ameaça e efeito de composição quando múltiplas operações privatizadas são aplicadas ao longo do tempo, consumindo um orçamento global [7, 36].

Para além de métricas formais, a literatura recorre a métricas empíricas de risco de divulgação que aproximam, em cenário experimental, a capacidade de um atacante inferir informação a partir de um modelo ou de um conjunto de dados privatizado. Um exemplo amplamente estudado são os ataques de *membership inference*, nos quais um atacante tenta decidir se um registo pertenceu ao conjunto de treino. A taxa de sucesso destes ataques (por exemplo, exatidão ou AUC do atacante acima do acaso) é usada como indicador do grau de exposição de informação individual [22, 24, 26, 70]. Uma extensão desta abordagem são os ataques baseados em exemplos canário, nos quais exemplos artificiais com combinações de valores improváveis são inseridos no conjunto de treino, e a diferença entre a perda do modelo nesses exemplos e a perda nos dados normais de treino é usada como indicador de memorização individual [27]. Outros trabalhos analisam ataques de inversão e de inferência de atributos, avaliando a capacidade de recuperar informação sensível a partir das saídas do modelo [26, 71]. Embora estas métricas não substituam garantias formais como DP, são úteis para comparar empiricamente abordagens em contextos específicos.

Do ponto de vista da utilidade, as métricas tendem a ser as mesmas usadas em cenários não privados. Em avaliação de risco de crédito, a AUC-ROC e o KS são particularmente relevantes por se relacionarem com medidas de discriminação amplamente usadas na prática bancária [1, 72]. Em contextos com desbalanceamento acentuado da variável alvo, o G-Mean (*Geometric Mean*) é uma métrica complementar que penaliza igualmente erros nas duas classes, sendo especialmente informativa quando um modelo que classifique todos os registos como a classe maioritária obteria métricas convencionais aparentemente elevadas [73]. Em muitos casos, considera-se também calibração (por exemplo, *Brier score* [74]) e estabilidade temporal, dado que modelos em produção, sobretudo em contexto regulado, não devem apenas discriminar bem, mas também manter um comportamento consistente ao longo do tempo. Em estudos de PPML, o objetivo é compreender como estas métricas se degradam à medida que aumenta o nível de proteção (por exemplo, ao reduzir ε ou intensificar perturbações em mecanismos locais).

Uma terceira dimensão é o custo computacional e operacional. Em PPML, este custo pode ser caracterizado por métricas como: (i) tempo de treino e de inferência, (ii) consumo de recursos de

computação, (iii) utilização de memória e armazenamento, (iv) volume de comunicação e número de rondas em cenários distribuídos, (v) escalabilidade com o número de registos e atributos e com o número de participantes, e (vi) complexidade de integração e de manutenção. Técnicas ao nível dos dados (por exemplo, anonimização, generalização ou LDP) tendem a ter custos moderados e podem ser aplicadas como pré-processamento, embora possam exigir reprocessamento se políticas ou requisitos mudarem. Em contraste, treino com mecanismos formais de privacidade (por exemplo, *DP-SGD*) e abordagens arquiteturais (por exemplo, aprendizagem federada com agregação segura ou computação segura) implicam frequentemente aumento do tempo de treino, consumo de memória, sobrecarga de comunicação e maior complexidade de engenharia e operação [7, 36].

Em síntese, a análise de compromissos em PPML assenta na combinação de: (i) métricas de privacidade (formais, quando existem, ou empíricas), (ii) métricas empíricas de risco de divulgação (por exemplo, sucesso de *membership inference*, *Attribute Inference* ou ataques baseados em exemplos canário), (iii) métricas de desempenho e qualidade do modelo e (iv) métricas de custo computacional e de integração. No contexto desta dissertação, a comparação entre abordagens de PPML em avaliação de risco de crédito será estruturada em torno destes três vértices — privacidade, desempenho e custo — para avaliar sistematicamente os compromissos associados às técnicas selecionadas.

2.3 Trabalho relacionado

A aplicação de técnicas de *Privacy-Preserving Machine Learning* (PPML) em contexto financeiro, e em particular em problemas de avaliação de risco de crédito, tem vindo a ganhar atenção nos últimos anos. Ainda assim, a literatura permanece relativamente fragmentada, com propostas frequentemente avaliadas em cenários e pressupostos técnicos específicos, organizando-se sobretudo em três linhas: métodos criptográficos, abordagens colaborativas e distribuídas, e técnicas de transformação ou perturbação dos dados de entrada.

Uma linha de trabalho centra-se em métodos criptográficos que viabilizam o cálculo de risco de crédito sobre dados encriptados. Divakar Allavarpu, Naresh e Krishna Mohan [75] propõem um enquadramento baseado em regressão logística com encriptação homomórfica, permitindo treinar e inferir na nuvem sem descodificar os atributos dos clientes, e reportam resultados competitivos com impacto limitado no desempenho para este tipo de modelo. De forma complementar, Andolfo et al. [76] exploram encriptação funcional aplicada a avaliação de risco de crédito, permitindo que a entidade que executa o cálculo obtenha apenas o resultado (por exemplo, uma probabilidade de incumprimento), sem acesso aos atributos em claro, discutindo o impacto no desempenho e o sobrecusto operacional da abordagem. Em conjunto, estes trabalhos mostram que é possível compatibilizar requisitos fortes de confidencialidade com modelos relativamente simples, tipicamente à custa de maior complexidade criptográfica e custo computacional.

Outra linha de investigação aborda cenários colaborativos e distribuídos, em que múltiplas instituições procuram beneficiar de informação multi-entidade sem partilhar diretamente dados

brutos. He et al. [77] propõem um método descentralizado de avaliação de risco de crédito baseado em informação multi-entidade, combinando mecanismos de partilha segura e integração de modelos locais para construir um modelo global sem centralizar dados sensíveis. Este tipo de abordagem é particularmente relevante quando a centralização de dados é limitada por restrições legais, concorrenciais ou de governação.

Mais recentemente, surgem propostas que atuam diretamente ao nível dos dados de entrada do modelo, através de transformação ou perturbação controlada, especialmente em cenários de aprendizagem automática como serviço. Prodomo et al. [78] apresentam uma técnica de ofuscação para avaliação de risco de crédito em que os dados são transformados antes de serem enviados para o prestador, procurando preservar a estrutura estatística útil ao modelo e reduzir o risco de divulgação de informação. Os autores avaliam o impacto no desempenho preditivo e na resistência a ataques de reconstrução. Esta linha é particularmente alinhada com o tipo de compromisso utilidade-privacidade discutido anteriormente, sendo a mais próxima da abordagem seguida nesta dissertação.

Para além de propostas especificamente focadas em avaliação de risco de crédito, existe literatura mais geral sobre PPML que recorre a variantes de encriptação homomórfica e computação multipartidária segura [68, 79]. No que diz respeito à perturbação não uniforme dos dados, a literatura documenta propostas de perturbação anisótropa guiada pela importância das variáveis, atribuindo menos ruído às variáveis com maior contributo preditivo para preservar a utilidade do modelo [80], e enquadramentos que ajustam a proteção de privacidade em função da sensibilidade de cada variável. Contudo, estas abordagens calibram a intensidade da perturbação pelo contributo preditivo ou pela sensibilidade estatística das variáveis, e não pela fuga de informação efetiva medida empiricamente. Muitos estudos permanecem ainda centrados numa técnica e num modelo concretos, o que dificulta comparações diretas entre famílias de métodos e a extrapolação para fluxos de processamento industriais.

Neste contexto, a presente dissertação diferencia-se por estruturar uma análise comparativa e sistemática de várias famílias de PPML aplicadas a avaliação de risco de crédito com dados tabulares e fluxo de processamento realista, avaliando não apenas a viabilidade técnica de cada abordagem, mas também os compromissos entre privacidade, desempenho preditivo e integração prática em sistemas bancários. Em particular, embora existam propostas de perturbação não uniforme na literatura, a calibração da intensidade da perturbação pela fuga de informação efetiva medida por *Attribute Inference* em dados tabulares de avaliação de risco de crédito não estava documentada, constituindo uma contribuição específica deste trabalho.

2.4 Conclusões

A revisão da literatura realizada ao longo deste capítulo permite sintetizar o estado da arte em avaliação de risco de crédito com *Machine Learning* e em *Privacy-Preserving Machine Learning* (PPML), bem como delimitar oportunidades de investigação particularmente relevantes para o contexto desta dissertação.

Em primeiro lugar, a área de avaliação de risco de crédito com dados tabulares encontra-se relativamente madura. A utilização de modelos estatísticos e de *Machine Learning* para estimar probabilidades de incumprimento é bem estabelecida, com metodologias clássicas baseadas em regressão logística e tabelas de pontuação de crédito, e com a adoção de modelos não lineares como *gradient boosting*, *random forests* ou *support vector machines* em estudos comparativos [1, 2, 81]. A literatura enfatiza também práticas específicas do domínio, como o tratamento de desbalanceamento, o recurso a discretização de variáveis e o uso de métricas como AUC-ROC e KS em contexto operacional e regulado [1, 72]. No entanto, estes desenvolvimentos assumem, em grande medida, acesso direto a dados em claro, remetendo a privacidade para processos organizacionais ou para medidas de pseudonimização com garantias limitadas.

Em segundo lugar, a literatura em privacidade aplicada a *Machine Learning* mostra que a simples remoção de identificadores diretos e a aplicação de anonimização clássica são, frequentemente, insuficientes para mitigar riscos de reidentificação ou de inferência, sobretudo quando se treinam modelos sobre dados sensíveis. Conceitos como *Differential Privacy* e a evidência empírica de ataques de *membership inference*, *Attribute Inference* e inversão de modelo indicam que os próprios modelos podem funcionar como canais de fuga de informação, motivando uma abordagem de proteção de dados desde a conceção que considere explicitamente a privacidade ao nível dos dados, dos algoritmos e da arquitetura [22, 24, 26, 56, 71].

Em terceiro lugar, a área de PPML oferece um conjunto diversificado de técnicas que atuam em diferentes níveis: mecanismos aplicados aos dados (anonimização, generalização, perturbação, *Local Differential Privacy*, dados sintéticos privados), mecanismos aplicados ao treino e ao modelo (por exemplo, treino com *Differential Privacy* no gradiente) e abordagens arquiteturais como aprendizagem federada, computação multipartidária segura ou enclaves de execução confiável [6, 7, 56]. Cada família apresenta vantagens e limitações, bem como requisitos distintos de integração e impactos em termos de desempenho e custo computacional. Ainda assim, a discussão destes compromissos surge frequentemente em cenários experimentais genéricos, nem sempre alinhados com as especificidades da avaliação de risco de crédito bancário.

Quando se restringe o foco ao contexto de crédito e banca, em vez das aplicações genéricas de PPML discutidas nos parágrafos anteriores, observa-se um conjunto crescente, mas ainda limitado, de trabalhos que aplicam estas técnicas especificamente a problemas de risco de crédito. Vários estudos exploram encriptação homomórfica e encriptação funcional para permitir avaliação de risco de crédito sobre dados cifrados, outros propõem abordagens descentralizadas ou baseadas em aprendizagem federada, e alguns investigam técnicas de perturbação ou transformação de dados para avaliação de risco de crédito disponibilizada como serviço [75–78]. Estes contributos demonstram viabilidade técnica, mas tendem a focar-se numa única família de técnicas, em modelos específicos (por exemplo, regressão logística cifrada) ou em cenários simplificados, limitando a comparabilidade entre abordagens e a generalização para contextos práticos reais.

Da análise da literatura emergem várias limitações nos trabalhos existentes:

- evidência ainda limitada de avaliações comparativas, num mesmo cenário experimental de avaliação de risco de crédito tabular, que confrontem diferentes famílias de PPML e utilizem

métricas alinhadas com o setor bancário [6, 7];

- necessidade de maior aproximação a fluxos de processamento realistas de avaliação de risco de crédito (incluindo desbalanceamento, discretização e modelos não lineares usados na prática) e de análise de como e onde integrar PPML ao longo desse fluxo [2, 81];
- articulação ainda pouco sistemática entre métricas formais de privacidade, métricas empíricas de risco de ataque (por exemplo, *membership inference* e *Attribute Inference*) e métricas de desempenho, para caracterizar de forma operacional os compromissos entre privacidade, utilidade e custo [24, 26, 56];
- escassez de propostas de arquitetura integrada de PPML especificamente orientadas para avaliação de risco de crédito tabular, combinando técnicas promissoras numa visão coerente e plausível do ponto de vista operacional [6, 77];
- ausência de abordagens de perturbação calibradas pela fuga de informação efetiva medida empiricamente por *Attribute Inference* em dados tabulares de avaliação de risco de crédito, distinguindo-se das propostas existentes que calibram a perturbação pelo contributo preditivo ou pela sensibilidade estatística das variáveis [80].

A dissertação posiciona-se, assim, na interseção entre avaliação de risco de crédito e PPML, contribuindo com (i) uma avaliação empírica comparativa de abordagens de preservação de privacidade num cenário real de avaliação de risco de crédito, (ii) a síntese dessas evidências numa proposta de arquitetura integrada de PPML para risco de crédito com aplicabilidade potencial em contexto bancário, e (iii) uma abordagem de perturbação seletiva calibrada pela fuga de informação efetiva por *Attribute Inference*, não documentada na literatura para dados tabulares de crédito.

Capítulo 3

Dados e Metodologia

Este capítulo apresenta os dados utilizados nesta dissertação, o fluxo de processamento que lhes é aplicado e a metodologia experimental que sustenta os experimentos descritos no capítulo seguinte. A caracterização do conjunto de dados e do seu contexto de origem é seguida pela descrição das etapas pelas quais os dados passam antes de chegarem aos modelos, identificando claramente as fases sobre as quais a investigação atua (discretização, codificação e treino do modelo) e as que ficam fora do âmbito desta dissertação, como a recolha e limpeza inicial dos dados. A definição dos modelos de base e das métricas de avaliação encerra o capítulo, estabelecendo os termos em que a utilidade e a privacidade são medidas ao longo de toda a experimentação descrita no Capítulo 4.

3.1 Origem e Caracterização dos Dados

Os dados utilizados nesta dissertação foram disponibilizados pela ITSCredit, empresa especializada em soluções de informação e análise de risco de crédito para o sector bancário, que desenvolve e mantém modelos de avaliação de risco de crédito utilizados por instituições financeiras no apoio a decisões de concessão de crédito. Trata-se de dados reais de operações de crédito, recolhidos no âmbito da atividade operacional de uma instituição financeira cliente da ITSCredit, o que permite que esta investigação seja conduzida num cenário real do sector, com as características, complexidades e restrições próprias de um sistema de avaliação de risco de crédito em produção.

Este contexto é particularmente relevante para a investigação em *Privacy-Preserving Machine Learning*. Por um lado, a informação financeira dos clientes é, pela sua própria natureza, especialmente sensível, expondo aspetos detalhados da situação económica e do comportamento de cada indivíduo. Por outro lado, o sector bancário está sujeito a exigências regulatórias específicas que reforçam a necessidade de proteção destes dados. Em conjunto, estes dois fatores colocam desafios concretos à aplicação de técnicas de preservação de privacidade em sistemas de decisão de crédito.

O conjunto de dados é composto por 59 435 registos, divididos em 41 604 para treino e 17 831 para teste. A variável alvo é binária, indicando a ocorrência ou não de incumprimento por parte do cliente, com uma taxa de incumprimento de aproximadamente 4,4%, um desbalanceamento característico de carteiras de crédito em contexto real, que condiciona tanto a seleção de métricas de avaliação como, como se verá ao longo do Capítulo 4, a aplicabilidade de diversas técnicas de PPML.

As variáveis disponíveis integram informação demográfica, comportamental e financeira dos clientes. Em conformidade com as práticas do domínio, todas são sujeitas a um processo de discretização, pelo qual atributos contínuos e categóricos são transformados em variáveis discretas com intervalos definidos, prática amplamente utilizada na construção de tabelas de pontuação de crédito, que contribui para a estabilidade e interpretabilidade dos modelos e facilita a validação e auditoria em contexto regulado.

Para o treino dos modelos foram selecionadas 14 variáveis com poder discriminativo relevante. Um número reduzido de variáveis facilita a interpretabilidade do modelo, permitindo compreender com clareza o contributo de cada uma para a decisão final, um requisito particularmente relevante em ambiente regulado. Reduz-se ainda o risco de incluir variáveis com pouco ou nenhum contributo preditivo, que apenas introduziriam ruído.

3.2 Fluxo de Processamento

O fluxo de processamento aplicado neste trabalho reflete as práticas operacionais utilizadas pela ITSCredit, alinhadas com as práticas estabelecidas no domínio da avaliação de risco de crédito, estando representado na Figura 3.1. A investigação atua sobre as etapas de discretização, codificação e treino do modelo; a recolha e limpeza inicial dos dados e a etapa final de pontuação de crédito ficam fora do âmbito desta dissertação, por refletirem decisões de negócio que não são objeto de estudo.

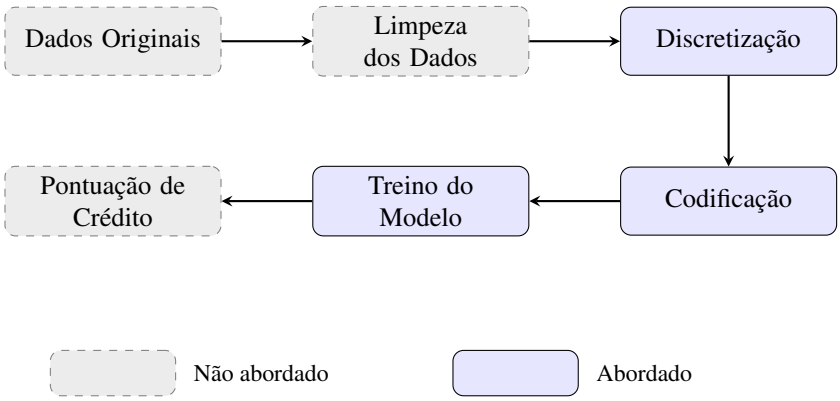

Figura 3.1: Fluxo de processamento de avaliação de risco de crédito utilizado neste trabalho

A etapa de discretização ocupa um lugar central nesta investigação, e não apenas como uma fase do fluxo de processamento. Em conformidade com as práticas do sector, todas as variáveis (numéricas e categóricas) são transformadas em variáveis discretas com intervalos definidos, processo amplamente utilizado na construção de tabelas de pontuação de crédito por motivos de estabilidade, interpretabilidade e auditoria em contexto regulado. Esta exigência de negócio, anterior e independente de qualquer consideração de privacidade, tem no entanto duas consequências diretas para a aplicação de *Privacy-Preserving Machine Learning* neste trabalho.

A primeira é que a própria discretização introduz uma forma indireta de proteção de privacidade: ao reduzir a granularidade dos dados individuais, dificulta a distinção entre registos que, nos dados originais, seriam distinguíveis. Esta proteção implícita deve ser tida em conta na interpretação dos resultados de privacidade ao longo da dissertação, já que parte da redução de risco observada pode não ser atribuível aos métodos de PPML aplicados, mas à própria estrutura dos dados.

A segunda consequência é mais determinante para a metodologia experimental: a discretização condiciona o ponto em que a perturbação pode ser aplicada, e cada opção traz uma limitação distinta. Aplicar a perturbação depois da discretização, quando as variáveis são já inteiramente categóricas, restringe os métodos de PPML disponíveis a mecanismos desenhados para dados categóricos, e exclui mecanismos formulados para variáveis contínuas. Aplicar a perturbação antes da discretização, sobre os dados originais com variáveis contínuas e categóricas, permite recorrer a um conjunto mais amplo de mecanismos, mas expõe o ruído introduzido ao risco de ser absorvido ou distorcido pelo processo de discretização subsequente, dificultando o controlo efetivo do nível de privacidade alcançado.

Esta tensão entre os dois pontos de aplicação da perturbação (antes ou depois da discretização) percorre toda a experimentação descrita no Capítulo 4, e resulta diretamente das características desta etapa do fluxo de processamento, que não pode ser alterada por se tratar de uma prática de negócio já estabelecida. Concluída a discretização, as variáveis são codificadas em formato binário multivariado (*one-hot encoding*) para serem fornecidas como entrada aos modelos, completando-se assim a preparação dos dados antes do treino.

3.3 Modelos de Base

Neste trabalho são utilizados dois modelos de classificação, selecionados de forma a cobrir o espectro habitual de abordagens em avaliação do risco de crédito: Regressão Logística e *Categorical Boosting* (*CatBoost*).

A Regressão Logística constitui o modelo de referência tradicional neste domínio. A sua utilização generalizada em sistemas de pontuação de crédito deve-se à interpretabilidade dos coeficientes, à compatibilidade com práticas de validação e auditoria em contexto regulado e à estabilidade do seu comportamento em dados tabulares. A sua inclusão neste trabalho permite avaliar os métodos de PPML no contexto do modelo mais amplamente utilizado na indústria.

O *CatBoost*, por sua vez, representa a família de modelos de gradient boosting, que consistentemente apresentam os melhores resultados em problemas de classificação com dados tabulares. A escolha deste modelo permite avaliar o comportamento das técnicas de privacidade num modelo de maior capacidade e desempenho preditivo superior.

Ambos os modelos foram treinados com configurações simples, sem otimização extensiva de hiperparâmetros, de forma a isolar o efeito dos métodos de PPML sobre o desempenho dos modelos sem o confundir com ganhos resultantes de afinação fina dos parâmetros. Em ambos os modelos foi ativado o mecanismo de ponderação de classes (*class weighting*), dado o desbalanceamento acentuado da variável alvo descrito na Secção 3.1: este mecanismo atribui um peso maior aos exemplos da classe minoritária, os casos de incumprimento, durante o treino, compensando a sua sub-representação e evitando que o modelo aprenda simplesmente a prever sempre a classe maioritária, o que resultaria numa elevada taxa de acerto aparente mas numa incapacidade prática de identificar clientes em risco. A utilização de dois modelos distintos permite ainda analisar se o impacto das técnicas de PPML é consistente entre abordagens de diferente complexidade.

3.4 Métricas de Avaliação de Utilidade

A avaliação da utilidade dos modelos assenta em métricas amplamente utilizadas no domínio da avaliação de risco de crédito, complementadas por uma métrica adicional motivada pelo desbalanceamento acentuado da variável alvo.

A AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*) mede a capacidade discriminativa global do modelo, quantificando a probabilidade de o modelo atribuir uma probabilidade de incumprimento mais elevada a um cliente que efetivamente entra em incumprimento do que a um cliente que não entra. É a métrica de referência no sector financeiro para comparação entre modelos de pontuação de crédito, sendo independente do limiar de classificação utilizado. Formalmente, a AUC-ROC é definida como:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{3.1}$$

onde *TPR* (*True Positive Rate*) e *FPR* (*False Positive Rate*) são calculados para cada limiar de classificação t.

A estatística KS (Kolmogorov-Smirnov) quantifica a distância máxima entre as distribuições acumuladas das probabilidades previstas para as duas classes. Na avaliação de risco de crédito, esta métrica é particularmente relevante por refletir diretamente a capacidade do modelo em separar clientes de risco elevado de clientes de baixo risco, sendo amplamente utilizada em contexto operacional e regulatório. O KS é definido como:

$$\text{KS} = \max_t |F_1(t) - F_0(t)| \tag{3.2}$$

onde $F_1(t)$ e $F_0(t)$ representam as funções de distribuição acumulada das probabilidades previstas para a classe positiva (incumprimento) e negativa (não-incumprimento), respetivamente.

Dado o desbalanceamento acentuado da variável alvo, com apenas 4,4% de registos positivos, as métricas tradicionais podem não capturar adequadamente o comportamento do modelo nas duas classes em simultâneo. Para complementar esta análise, é utilizado o *G-Mean* (*Geometric Mean*), definido como a raiz quadrada do produto entre a *Sensitivity* e a *Specificity*:

$$\text{G-Mean} = \sqrt{\textit{Sensitivity} \times \textit{Specificity}} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{3.3}$$

onde TP, TN, FP e FN representam os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respetivamente. Esta métrica penaliza igualmente erros nas duas classes, sendo especialmente informativa em cenários de desbalanceamento, uma vez que um modelo que classifique todos os registos como não-incumprimento obteria um *G-Mean* nulo, ao contrário do que sucederia com métricas como a *accuracy*.

3.5 Métricas de Avaliação de Privacidade

A avaliação de privacidade dos modelos é realizada através de três métricas empíricas, selecionadas por medirem diretamente as ameaças de privacidade mais relevantes no contexto da avaliação de risco de crédito.

Membership Inference Attack (MIA)

O MIA avalia a capacidade de um atacante em determinar se um determinado registo pertenceu ao conjunto de treino do modelo, com base apenas nas probabilidades de saída [24]. Um modelo com MIA próximo de 0.5 é resistente a este ataque, pois o atacante não consegue distinguir membros de não-membros melhor do que o acaso. Valores superiores indicam que o modelo expõe informação sobre os dados de treino.

O MIA é expresso como a AUC-ROC do classificador do atacante, assumindo valores no intervalo $[0,1]$ [26]. A interpretação do valor obtido é a seguinte:

- **0.5** — o atacante não consegue distinguir membros de não-membros melhor do que o acaso, indicando ausência de fuga de informação detetável;
- **Entre 0.5 e 1.0** — quanto mais o valor se afasta de 0.5 em direção a 1.0, maior a capacidade do atacante em identificar registos que pertenceram ao conjunto de treino, e maior a exposição de informação por parte do modelo;
- **1.0** — o atacante consegue distinguir perfeitamente membros de não-membros, correspondendo à máxima fuga de informação possível.

Esta métrica foi escolhida por ser a medida empírica de *membership inference* mais amplamente utilizada na literatura de PPML [22], permitindo comparação direta com trabalhos relacionados.

Attribute Inference Attack

O *Attribute Inference Attack* mede a capacidade de um atacante em inferir o valor de uma variável sensível de um indivíduo a partir das probabilidades de saída do modelo [26]. O ataque funciona da seguinte forma: o atacante tem conhecimento dos valores de todas as restantes variáveis do indivíduo e utiliza as probabilidades de saída do modelo para tentar adivinhar o valor da variável que desconhece. Para quantificar o risco real introduzido pelo modelo, o resultado do ataque é comparado com uma referência, nomeadamente o que o atacante conseguiria inferir usando apenas as restantes variáveis, sem qualquer acesso ao modelo. O ganho sobre esta referência mede a informação adicional que o modelo expõe sobre a variável sensível.

O ganho assume valores no intervalo $[0, 1]$. A interpretação do valor obtido é a seguinte:

- **0** — o modelo não acrescenta informação além do que já seria dedutível pelas restantes variáveis, indicando ausência de fuga de informação adicional;
- **Entre 0 e 1** — quanto maior o valor, maior a informação adicional que o modelo expõe sobre a variável sensível, refletindo uma maior fuga de informação individual;
- **1** — o modelo revela completamente o valor da variável sensível, correspondendo à máxima fuga de informação possível.

Esta métrica foi escolhida por medir diretamente a fuga de informação sobre variáveis individuais sensíveis, uma ameaça particularmente relevante em contextos onde os dados integram informação de natureza demográfica ou financeira, tipicamente presentes em sistemas de avaliação de risco de crédito.

Canary Gap

O Canary Gap mede a memorização de exemplos específicos pelo modelo [27]. São inseridos no conjunto de treino exemplos artificiais (os canários) com combinações de valores improváveis, e avalia-se a diferença entre a função de perda do modelo nesses exemplos e a função de perda nos dados normais de treino [27]. Ao contrário do MIA e do *Attribute Inference Attack*, o Canary Gap não está confinado ao intervalo $[0, 1]$, podendo assumir qualquer valor real. A interpretação do valor obtido é a seguinte:

- **Valor positivo elevado** — o modelo atribui aos canários uma perda muito superior à dos dados normais de treino, indicando que não os memorizou e os trata como exemplos desconhecidos;
- **Valor próximo de zero** — o modelo atribui aos canários uma perda semelhante à dos dados de treino, sugerindo algum grau de memorização;
- **Valor negativo** — o modelo atribui aos canários uma perda inferior à dos dados normais de treino, indicando memorização direta dos exemplos artificiais.

Esta métrica foi escolhida por complementar o MIA na avaliação de memorização: enquanto o MIA avalia separabilidade estatística entre treino e teste, o Canary Gap avalia diretamente se o modelo retém informação sobre exemplos individuais específicos [27]. É de notar que, em dados inteiramente categóricos como os deste trabalho, os canários não podem ser verdadeiramente impossíveis. Ao contrário do que sucede com variáveis contínuas, onde é possível inserir valores fora de qualquer intervalo plausível, em variáveis categóricas apenas é possível construir combinações de categorias existentes que sejam improváveis. Isto significa que os canários podem, em teoria, corresponder a registos reais, o que reduz a nitidez do sinal de memorização. Trata-se de uma limitação reconhecida da métrica quando aplicada a dados categóricos [27], que não invalida a sua utilização mas deve ser tida em conta na interpretação dos resultados.

Capítulo 4

Análise Experimental

Este capítulo apresenta as experiências realizadas no âmbito desta dissertação, estruturadas de forma a refletir a progressão metodológica seguida ao longo do trabalho. O ponto de partida é o modelo real, treinado sem qualquer mecanismo de preservação de privacidade, que serve de referência para todas as comparações subsequentes.

A investigação foi conduzida de forma iterativa, em que os resultados de cada conjunto de experiências motivaram as decisões metodológicas seguintes. São avaliadas técnicas pertencentes a diferentes famílias de PPML (métodos com garantia formal de privacidade, métodos de síntese de dados, métodos de proteção ao nível do modelo e métodos de perturbação direta dos dados), culminando numa abordagem de perturbação seletiva desenvolvida iterativamente com base nos resultados obtidos ao longo da experimentação.

Para cada método são apresentadas as métricas de utilidade e de privacidade definidas no Capítulo 3, permitindo uma análise comparativa sistemática dos compromissos entre privacidade e desempenho ao longo de todas as abordagens testadas.

4.1 Modelo de Referência

O modelo real, treinado sem qualquer mecanismo de preservação de privacidade, serve de referência para todas as comparações realizadas ao longo deste capítulo. Os resultados obtidos estabelecem o ponto de partida em termos de utilidade e de exposição de privacidade, permitindo quantificar o impacto de cada técnica de PPML relativamente ao modelo original.

Métricas de Utilidade e Privacidade Global

A Tabela 4.1 apresenta as métricas de utilidade e de privacidade global para os dois modelos de base. Ambos os modelos apresentam desempenho preditivo semelhante, com AUC-ROC de 0.779 e 0.781 para a Regressão Logística e o *CatBoost*, respetivamente. Em termos de privacidade, o MIA próximo de 0.50 em ambos os modelos indica resistência ao ataque de *membership inference*. O Canary Gap revela, no entanto, um comportamento distinto entre os dois modelos: a Regressão Logística apresenta um gap de 19.49, indicando forte resistência à memorização de exemplos

individuais, enquanto o *CatBoost* apresenta um gap de 1.90, sugerindo uma maior tendência para memorizar exemplos específicos do conjunto de treino.

Tabela 4.1: Métricas de utilidade e privacidade global — Modelo de Referência

Modelo	AUC-ROC	KS	G-Mean	MIA	Canary Gap
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90

LR (Ref.) — Regressão Logística sem mecanismo de privacidade; CAT (Ref.) — *CatBoost* sem mecanismo de privacidade.

Attribute Inference

A Tabela 4.2 apresenta os resultados do ataque de *Attribute Inference* para cada variável, nos dois modelos. O valor reportado corresponde ao ganho sobre a referência, ou seja, a informação adicional que o modelo expõe sobre cada variável, relativamente ao que seria possível inferir sem acesso ao modelo, conforme definido na Secção 3.5.

Tabela 4.2: *Attribute Inference* — Modelo de Referência

Variável	LR (Ref.)	CAT (Ref.)
Profissao	0.2065	0.2508
TotOutflowU3	0.1566	0.2312
Prazo	0.1075	0.1480
Idade	0.1187	0.1576
TipoCC	0.1071	0.1032
Finalidade	0.0871	0.1077
AntBanco	0.1221	0.1241
DifNumCredSFDtDt3	0.0600	0.0822
RacMontanteExpTot	0.0680	0.0688
ValorVencMedioSFU12	0.0297	0.0421
ExpSFDt6	0.0075	0.0060
NumCredSFDt6	0.0228	0.0092
RacSaldoMedioMinU6	0.0030	0.0034
FlagVencU12	0.0000	0.0000

Os resultados revelam um padrão consistente entre os dois modelos. As variáveis com maior fuga de informação são a Profissao e o TotOutflowU3, com ganhos de 0.207 e 0.157 na Regressão Logística e de 0.251 e 0.231 no *CatBoost*, respetivamente. Estas variáveis são também as mais preditivas do modelo, o que sugere que a fuga de informação é estrutural, proporcional ao poder discriminativo de cada variável. No extremo oposto, variáveis como RacSaldoMedioMinU6 e FlagVencU12 apresentam fuga de informação negligenciável, refletindo o seu reduzido contributo para o modelo.

4.2 Métodos de Privacidade Diferencial Local

Os métodos de *Local Differential Privacy* (LDP) inserem-se na família de técnicas de PPML ao nível dos dados, descrita na Secção 2.2.1. Nestes métodos, a perturbação é aplicada diretamente aos dados de cada registo individual antes do treino do modelo, garantindo que a contribuição individual não pode ser distinguida com uma probabilidade superior a e^{ε}, onde ε é o parâmetro de privacidade que controla o nível de proteção. Valores baixos de ε correspondem a maior privacidade mas também a maior ruído introduzido nos dados, enquanto valores elevados se aproximam do comportamento sem perturbação.

A escolha destes métodos como primeira família a avaliar justifica-se pelo seu rigor teórico, pois oferecem garantias formais de privacidade, independentemente do modelo utilizado e das características dos dados [45], e pela sua relevância na literatura de PPML. Ao aplicar a perturbação ao nível do registo individual, estes métodos são também naturalmente compatíveis com o fluxo de processamento de avaliação de risco de crédito utilizado, não exigindo alterações ao processo de treino do modelo.

Nesta secção são avaliados três mecanismos LDP: o *Generalized Randomized Response* (GRR) [47], o *Optimized Unary Encoding* (OUE) [82] e o *Randomized Aggregatable Privacy-Preserving Ordinal Response* (RAPPOR) [34]. O GRR e o RAPPOR são aplicados às variáveis categóricas antes da codificação binária multivariada, enquanto o OUE opera diretamente sobre a representação binária após a codificação. Os três métodos são avaliados com diferentes níveis de privacidade definidos pelo parâmetro ε.

4.2.1 GRR — *Generalized Randomized Response*

O *Generalized Randomized Response* (GRR) foi o primeiro mecanismo de privacidade diferencial local a ser avaliado neste trabalho, por constituir a abordagem mais direta e teoricamente fundamentada para privatizar variáveis categóricas com garantia formal. O GRR é um mecanismo clássico de *randomized response* generalizado para domínios com múltiplas categorias [47], oferecendo garantia ε-LDP com uma formulação matemática simples e interpretável

Para cada valor verdadeiro de uma variável com k categorias, o mecanismo mantém o valor original com probabilidade:

$$p = \frac{e^{\varepsilon}}{e^{\varepsilon} + k - 1} \tag{4.1}$$

e substitui por uma categoria aleatória com probabilidade:

$$q = \frac{1}{e^{\varepsilon} + k - 1} \tag{4.2}$$

onde ε é o parâmetro de privacidade. Valores mais baixos de ε correspondem a maior privacidade: com $\varepsilon \to 0$ o mecanismo responde aleatoriamente independentemente do valor verdadeiro, e com $\varepsilon \to \infty$ o valor verdadeiro é sempre reportado sem perturbação.

O GRR é aplicado às 14 variáveis categóricas diretamente sobre os dados discretizados, antes da codificação binária multivariada. Esta escolha de ponto de aplicação é natural para este mecanismo, uma vez que o GRR opera sobre categorias discretas, exatamente a representação que resulta da discretização. A perturbação é aplicada independentemente a cada variável e a cada registo do conjunto de treino, mantendo o conjunto de teste original inalterado para avaliação.

Foram testados três níveis de privacidade, escolhidos de forma a cobrir o espectro entre proteção forte e fraca neste contexto:

- **Forte** ($\varepsilon = 0.5$) — perturbação elevada, com probabilidade de manter o valor original próximo de $\frac{1}{k}$ para variáveis com muitas categorias;
- **Médio** ($\varepsilon = 1.0$) — compromisso entre privacidade e utilidade, dentro do intervalo tipicamente explorado na literatura de LDP [45];
- **Fraco** ($\varepsilon = 2.0$) — perturbação moderada, próximo do comportamento sem perturbação para variáveis com poucas categorias.

É importante notar que o efeito prático de cada valor de ε depende do número de categorias de cada variável: para variáveis com muitas categorias como a Profissao (8 categorias), mesmo $\varepsilon = 2.0$ introduz perturbação significativa, enquanto para variáveis binárias o impacto é menor.

A Tabela 4.3 apresenta as métricas de utilidade e privacidade global para os três níveis de privacidade.

Tabela 4.3: Métricas de utilidade e privacidade global — GRR

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
GRR Forte LR (ε=0.5)	0.6957	0.2981	0.6183	0.7188	0.26
GRR Médio LR (ε=1.0)	0.7437	0.3834	0.6797	0.7056	1.52
GRR Fraco LR (ε=2.0)	0.7615	0.4059	0.6978	0.6571	3.85
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
GRR Forte CAT (ε=0.5)	0.6571	0.2327	0.5992	0.6948	-0.28
GRR Médio CAT (ε=1.0)	0.7322	0.3582	0.6645	0.6928	0.32
GRR Fraco CAT (ε=2.0)	0.7570	0.3987	0.6863	0.6529	1.27

Os resultados evidenciam um compromisso claro entre privacidade e utilidade. Com $\varepsilon = 0.5$, a AUC-ROC desce para 0.696 no LR e 0.657 no *CatBoost*, uma degradação significativa que torna estes modelos pouco adequados para uso em produção. À medida que o epsilon aumenta para 2.0, a utilidade recupera parcialmente, mas nunca atinge os valores do modelo real, confirmando que qualquer nível de perturbação GRR tem um custo em desempenho.

O MIA sobe em todos os níveis de perturbação, atingindo 0.719 no GRR Forte LR. O atacante consegue distinguir os dados de treino perturbados dos dados de teste originais com base na diferença estatística entre as duas distribuições, efeito do desfasamento introduzido pela perturbação, e não da memorização individual de exemplos pelo modelo.

O Canary Gap do GRR Forte *CatBoost* é negativo (-0.28), o que indica que o modelo memoriza os canários, resultado que se explica pela perturbação excessiva: com muito ruído nos dados de treino, o modelo perde capacidade de generalização e tende a memorizar os poucos exemplos que consegue aprender, incluindo os canários artificiais.

Tabela 4.4: *Attribute Inference* — GRR (variáveis de maior fuga de informação)

Variável	**LR (Ref.)**	**GRR Forte LR**	**GRR Médio LR**	**GRR Fraco LR**
Profissao	0.2065	0.1716	0.1770	0.2013
TotOutflowU3	0.1566	0.3148	0.1374	0.2095
Prazo	0.1075	0.0604	0.0708	0.1084
Idade	0.1187	0.2351	0.1877	0.1550
Variável	**CAT (Ref.)**	**GRR Forte CAT**	**GRR Médio CAT**	**GRR Fraco CAT**
Profissao	0.2508	0.1578	0.2084	0.2364
TotOutflowU3	0.2312	0.3430	0.1477	0.2436
Prazo	0.1480	0.0632	0.0574	0.1207
Idade	0.1576	0.2434	0.1929	0.1776

O *Attribute Inference* revela um comportamento inesperado. Seria de esperar que mais perturbação resultasse em menos fuga de informação, mas os resultados mostram o contrário em algumas variáveis. No GRR Forte, o TotOutflowU3 sobe de 0.157 para 0.315 no LR e de 0.231 para 0.343 no *CatBoost*, e a Idade de 0.119 para 0.235 no LR e de 0.158 para 0.243 no *CatBoost*. A perturbação uniforme de todas as variáveis revela um efeito consistente entre os dois modelos: ao reduzir o sinal preditivo das variáveis menos importantes, as variáveis mais fortemente correlacionadas com a variável alvo ficam proporcionalmente mais influentes nas previsões do modelo, tornando-as mais fáceis de inferir por um atacante. Este resultado levantou uma questão que se revelaria transversal a toda a investigação: a fuga de informação por inferência de atributos não parece ser simplesmente proporcional à quantidade de perturbação aplicada, mas depende de algo mais profundo na relação entre as variáveis e a variável alvo.

O GRR Médio apresenta o melhor compromisso em ambos os modelos (AUC-ROC de 0.744 no LR e 0.732 no *CatBoost*), com redução de Profissao para 0.177 e 0.208 e de TotOutflowU3 para 0.137 e 0.148, respetivamente. A degradação de utilidade face ao modelo de referência e o MIA elevado levantaram a questão de se aplicar a perturbação diretamente sobre a representação binária após a codificação produziria resultados diferentes. O OUE foi desenvolvido precisamente para este ponto de aplicação, motivando a sua avaliação na secção seguinte.

4.2.2 OUE — *Optimized Unary Encoding*

A questão levantada pelos resultados do GRR, nomeadamente se perturbar após a codificação produziria resultados diferentes, motivou a avaliação do OUE, mecanismo desenhado especificamente para operar sobre representações binárias de alta dimensionalidade. O *Optimized Unary Encoding* (OUE) é um mecanismo de privacidade diferencial local desenvolvido especificamente

para codificações binárias de alta dimensionalidade [82]. Ao contrário do GRR, que opera diretamente sobre as categorias originais, o OUE é aplicado após a codificação binária multivariada, perturbando cada bit do vetor binário de forma independente.

Seja $b \in \{0,1\}$ o valor verdadeiro de um bit do vetor codificado e $r \in \{0,1\}$ o valor reportado pelo mecanismo após perturbação. O OUE define as seguintes probabilidades condicionadas:

$$P(r=1 \mid b=1) = \frac{1}{2}, \quad P(r=1 \mid b=0) = \frac{1}{e^{\varepsilon}+1} \tag{4.3}$$

Esta assimetria entre as duas probabilidades condicionadas é a principal característica que distingue o OUE de mecanismos simétricos mais simples, e foi demonstrado ser ótima em termos de variância para estimação de frequências em domínios de alta cardinalidade [82].

O OUE é aplicado após a codificação binária multivariada das 14 variáveis, atuando diretamente sobre a representação binária utilizada como entrada pelos modelos. Tal como no GRR, são testados três níveis de privacidade (forte com $\varepsilon = 0.5$, médio com $\varepsilon = 1.0$ e fraco com $\varepsilon = 2.0$), permitindo comparação direta entre os dois mecanismos nas mesmas condições experimentais.

A Tabela 4.5 apresenta as métricas de utilidade e privacidade global.

Tabela 4.5: Métricas de utilidade e privacidade global — OUE

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
OUE Forte LR (ε=0.5)	0.6920	0.2885	0.6112	0.7019	-0.04
OUE Médio LR (ε=1.0)	0.7270	0.3407	0.6558	0.6936	0.10
OUE Fraco LR (ε=2.0)	0.7488	0.3741	0.6855	0.6787	0.58
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
OUE Forte CAT (ε=0.5)	0.6527	0.2395	0.5963	0.6853	0.03
OUE Médio CAT (ε=1.0)	0.7170	0.3280	0.6374	0.6930	0.17
OUE Fraco CAT (ε=2.0)	0.7470	0.3822	0.6849	0.6725	0.24

O padrão de resultados é similar ao GRR, com degradação crescente de utilidade com epsilon mais baixo e MIA elevado em todos os níveis por desfasamento de distribuição. O OUE Forte LR tem Canary Gap negativo (-0.04), indicando memorização, tal como aconteceu no GRR Forte.

Comparando diretamente com o GRR, o OUE apresenta resultados ligeiramente inferiores em utilidade para os mesmos níveis de epsilon: o OUE Fraco LR tem AUC-ROC de 0.749 face a 0.762 do GRR Fraco LR. Uma diferença relevante é o Canary Gap: enquanto o GRR Fraco LR tem gap de 3.85, o OUE Fraco LR tem apenas 0.58, sugerindo que o OUE, ao perturbar ao nível do bit após codificação, induz mais memorização do que o GRR que perturba ao nível da categoria original.

Tabela 4.6: *Attribute Inference* — OUE (variáveis de maior fuga de informação)

Variável	**LR (Ref.)**	**OUE Forte LR**	**OUE Médio LR**	**OUE Fraco LR**
Profissao	0.2065	0.1667	0.2069	0.2469
TotOutflowU3	0.1566	0.1778	0.1815	0.1938
Prazo	0.1075	0.0707	0.1060	0.1037
Idade	0.1187	0.1626	0.1942	0.1710
Variável	**CAT (Ref.)**	**OUE Forte CAT**	**OUE Médio CAT**	**OUE Fraco CAT**
Profissao	0.2508	0.1630	0.2245	0.2641
TotOutflowU3	0.2312	0.2708	0.2320	0.1839
Prazo	0.1480	0.1034	0.0940	0.0933
Idade	0.1576	0.1611	0.2217	0.2034

O *Attribute Inference* no OUE revela um padrão preocupante. O TotOutflowU3 mantém-se elevado em todos os níveis no LR (0.178, 0.182 e 0.194 para forte, médio e fraco) e no *CatBoost* o padrão é ainda menos favorável, com valores entre 0.184 e 0.271. A Profissao reduz ligeiramente no OUE Forte (0.167 no LR, 0.163 no *CatBoost*) mas sobe acima do modelo de referência no OUE Fraco (0.247 e 0.264, respetivamente). O OUE não demonstra vantagem sobre o GRR na fuga de informação por inferência de atributos, reforçando a questão levantada na secção anterior sobre a relação entre perturbação e fuga de informação estrutural neste conjunto de dados.

A comparação entre GRR e OUE mostrou que o ponto de aplicação da perturbação, antes ou depois da codificação, não resolve o problema estrutural da fuga de informação. Esta conclusão motivou a avaliação do RAPPOR, que introduz uma dupla camada de perturbação com o objetivo de reforçar as garantias de privacidade.

4.2.3 RAPPOR — *Randomized Aggregatable Privacy-Preserving Ordinal Response*

O *Randomized Aggregatable Privacy-Preserving Ordinal Response* (RAPPOR) é um mecanismo de privacidade diferencial local originalmente desenvolvido pela Google para recolha de estatísticas de utilização em larga escala [34]. A sua característica distintiva face ao GRR e ao OUE é a aplicação de duas camadas de perturbação sequenciais: uma *Permanent Randomized Response* (PRR), que gera uma máscara permanente associada ao valor verdadeiro, e uma *Instantiated Randomized Response* (IRR), que adiciona ruído adicional em cada reporte. Esta dupla camada visa dificultar a correlação entre múltiplos reportes do mesmo indivíduo ao longo do tempo, uma característica relevante no contexto original de recolha contínua de dados, mas que no presente trabalho, onde cada registo é reportado uma única vez, se traduz simplesmente numa perturbação mais agressiva para os mesmos valores de ε.

No contexto deste trabalho, o RAPPOR é aplicado às variáveis categóricas antes da codificação binária multivariada, tal como o GRR. A implementação segue a formulação original [34], com as probabilidades da IRR calibradas pelo parâmetro ε:

$$p = \frac{1}{e^{\varepsilon}+1}, \quad q = \frac{e^{\varepsilon}}{e^{\varepsilon}+1} \tag{4.4}$$

Foram testados cinco níveis de privacidade ($\varepsilon \in \{0.5, 1.0, 2.0, 5.0, 10.0\}$), de forma a cobrir um espectro mais amplo que o GRR e OUE, dado que a dupla camada de perturbação torna o RAPPOR mais agressivo para os mesmos valores de ε.

A Tabela 4.7 apresenta as métricas de utilidade e privacidade global.

Tabela 4.7: Métricas de utilidade e privacidade global — RAPPOR

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
RAPPOR LR (ε=0.5)	0.6239	0.2049	0.5935	0.6794	0.23
RAPPOR LR (ε=1.0)	0.7221	0.3513	0.6345	0.7311	0.89
RAPPOR LR (ε=2.0)	0.7582	0.4072	0.6451	0.7408	2.57
RAPPOR LR (ε=5.0)	0.7645	0.4264	0.6969	0.6887	4.62
RAPPOR LR (ε=10.0)	0.7657	0.4302	0.7006	0.6721	4.50
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
RAPPOR CAT (ε=0.5)	0.5449	0.0889	0.5232	0.6799	0.18
RAPPOR CAT (ε=1.0)	0.7039	0.3298	0.6370	0.7120	0.27
RAPPOR CAT (ε=2.0)	0.7370	0.3584	0.6297	0.7174	0.17
RAPPOR CAT (ε=5.0)	0.7545	0.4068	0.6662	0.7011	0.33
RAPPOR CAT (ε=10.0)	0.7599	0.4175	0.6923	0.6828	0.72

Os resultados do RAPPOR confirmam o padrão já observado no GRR e OUE, mas de forma mais acentuada. Com $\varepsilon = 0.5$, o *CatBoost* tem AUC-ROC de apenas 0.545, o valor mais baixo de todos os métodos testados, tornando o modelo completamente inutilizável. O LR é mais robusto mas ainda assim com AUC-ROC de 0.624, muito abaixo do aceitável para uso em produção.

Um resultado particularmente relevante é o comportamento do MIA com o RAPPOR. No GRR, o MIA diminuía progressivamente com epsilon mais alto; no RAPPOR mantém-se elevado mesmo com $\varepsilon = 10.0$ (0.672 no LR), sugerindo que a camada permanente de perturbação cria um desfasamento mais persistente entre treino e teste.

O Canary Gap do *CatBoost* mantém-se muito baixo em todos os níveis, entre 0.17 e 0.72, confirmando a tendência de memorização deste modelo com dados perturbados.

Tabela 4.8: *Attribute Inference* — RAPPOR (variáveis de maior fuga de informação)

Variável	**LR (Ref.)**	**RAPPOR LR (ε=0.5)**	**RAPPOR LR (ε=1.0)**	**RAPPOR LR (ε=2.0)**
Profissao	0.2065	0.1564	0.2092	0.1634
TotOutflowU3	0.1566	0.1624	0.1828	0.1587
Prazo	0.1075	0.0482	0.0460	0.0662
Idade	0.1187	0.2286	0.1353	0.1585
Variável	**CAT (Ref.)**	**RAPPOR CAT (ε=0.5)**	**RAPPOR CAT (ε=1.0)**	**RAPPOR CAT (ε=2.0)**
Profissao	0.2508	0.1593	0.2269	0.2236
TotOutflowU3	0.2312	0.2157	0.2239	0.2151
Prazo	0.1480	0.0434	0.0465	0.0673
Idade	0.1576	0.2521	0.1731	0.1955

O *Attribute Inference* no RAPPOR confirma o padrão já identificado no GRR e no OUE: a fuga de informação nas variáveis mais problemáticas mantém-se resistente à perturbação, independentemente do nível de epsilon ou do mecanismo utilizado. Com $\varepsilon = 0.5$, a Idade sobe de 0.119 para 0.229 no LR e de 0.158 para 0.252 no *CatBoost*. A Profissao melhora ligeiramente com epsilon baixo em ambos os modelos mas volta a subir com epsilon mais alto. O TotOutflowU3 mantém-se sem melhoria consistente em nenhum nível, com valores próximos ou acima do modelo de referência em ambos os modelos.

Comparando diretamente com o GRR para os mesmos valores de ε, o RAPPOR apresenta resultados inferiores em utilidade sem vantagem na fuga de informação das variáveis mais problemáticas, sugerindo que a dupla camada de perturbação, neste contexto de dados categóricos com uma única recolha por registo, não acrescenta valor face ao GRR.

Os três mecanismos LDP avaliados (GRR, OUE e RAPPOR) partilham uma limitação comum: a perturbação uniforme sobre todas as variáveis degrada a utilidade sem conseguir eliminar a fuga de informação nas variáveis estruturalmente mais problemáticas. Uma abordagem alternativa seria abandonar a perturbação direta dos dados originais e, em vez disso, gerar dados sintéticos que preservem as propriedades estatísticas do conjunto original com garantias formais de privacidade, substituindo os dados reais por dados artificiais que nunca expõem registos individuais. Esta família de métodos é avaliada na secção seguinte.

4.3 Dados Sintéticos com Privacidade Diferencial

Os métodos de geração de dados sintéticos com Privacidade Diferencial inserem-se na família de técnicas de PPML ao nível dos dados, descrita na Secção 2.2.1. A geração de dados sintéticos consiste na criação de um conjunto de dados artificial que reproduz as propriedades estatísticas do conjunto de dados original (distribuições marginais, correlações entre variáveis e relação com a variável alvo), sem conter registos reais de clientes. O modelo de aprendizagem automática é depois treinado exclusivamente sobre os dados sintéticos, sem acesso aos dados originais.

Quando combinada com Privacidade Diferencial, a geração de dados sintéticos garante formalmente que o processo de geração não revela informação sobre nenhum indivíduo específico do conjunto de dados original [51]. O parâmetro ε controla o nível de privacidade: valores baixos correspondem a dados sintéticos com menor fidelidade estatística, isto é, as distribuições e correlações geradas afastam-se mais das do conjunto de dados original, mas com maior proteção de privacidade; valores altos aproximam os dados sintéticos das propriedades estatísticas reais, com menor proteção. Uma abordagem concetualmente diferente dos mecanismos LDP: em vez de perturbar os dados de treino, são gerados dados artificiais que o modelo nunca viu, reduzindo o risco de memorização de exemplos individuais do conjunto de dados real [56].

Nesta secção são avaliados quatro métodos:

- **DP FT SD** (*Differentially Private Frequency Table Synthetic Data*) [50]
- **DP-CTGAN** (*Differentially Private Conditional Tabular GAN*) [36, 54]
- **CTGAN** (*Conditional Tabular GAN*) [54]
- **CTGAN+DPGAN** (*Conditional Tabular GAN* com *Differentially Private GAN*) [55]

4.3.1 DP FT SD — Dados Sintéticos por Tabelas de Frequências com Privacidade Diferencial

O DP FT SD é um método de geração de dados sintéticos baseado em tabelas de frequências com Privacidade Diferencial [50]. Para cada variável, é estimada a tabela de frequências das categorias com adição de ruído calibrado pelo parâmetro ε, garantindo que as frequências publicadas não revelam informação sobre nenhum indivíduo específico [50]. A geração de registos sintéticos é feita por amostragem a partir das distribuições privatizadas de cada variável, condicionada à classe da variável alvo.

Este método foi o primeiro a ser testado na família de dados sintéticos por ser o mais simples concetualmente: não requer treino de modelos generativos complexos e a sua implementação é direta. A independência entre variáveis na amostragem é simultaneamente a sua principal limitação: ao não modelar correlações entre variáveis, os dados sintéticos podem não preservar relações importantes para o modelo de avaliação de risco de crédito.

Foram testados três níveis de privacidade: forte ($\varepsilon = 0.5$), médio ($\varepsilon = 1.0$) e fraco ($\varepsilon = 2.0$). O modelo é treinado exclusivamente sobre os dados sintéticos gerados, sendo avaliado sobre os dados de teste reais.

A Tabela 4.9 apresenta as métricas de utilidade e privacidade global.

Tabela 4.9: Métricas de utilidade e privacidade global — DP FT SD

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
FT Forte LR	0.7512	0.4052	0.7013	0.6439	18.79
FT Médio LR	0.7662	0.4220	0.7074	0.6516	19.78
FT Fraco LR	0.7685	0.4275	0.7090	0.6532	21.53
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
FT Forte CAT	0.7505	0.3898	0.6849	0.6635	1.88
FT Médio CAT	0.7573	0.4001	0.6939	0.6695	3.07
FT Fraco CAT	0.7603	0.4062	0.7008	0.6512	3.13

Os resultados revelam um comportamento distinto face aos métodos LDP testados anteriormente. A degradação de utilidade é moderada: o FT Fraco LR atinge AUC-ROC de 0.769, apenas 0.010 abaixo do modelo de referência, e mesmo o FT Forte LR mantém 0.751. No *CatBoost* o padrão é semelhante, com AUC-ROC entre 0.751 e 0.760.

O Canary Gap mantém-se elevado em ambos os modelos: o LR tem gaps entre 18.8 e 21.5, e o *CatBoost* entre 1.9 e 3.1, valores próximos ou superiores ao modelo de referência. Um modelo treinado sobre dados sintéticos aprende distribuições estatísticas artificiais, o que reduz a sua capacidade de memorizar combinações de valores específicas presentes nos dados reais.

O MIA sobe para 0.64–0.67 em todos os modelos, refletindo o desfasamento entre os dados sintéticos usados no treino e os dados reais usados no teste, o mesmo fenómeno observado nos métodos LDP, mas agora amplificado pela natureza completamente artificial dos dados de treino.

A Tabela 4.10 apresenta os resultados do *Attribute Inference* para as variáveis de maior fuga de informação.

Tabela 4.10: *Attribute Inference* — DP FT SD (variáveis de maior fuga de informação)

Variável	**LR (Ref.)**	**FT Forte LR**	**FT Médio LR**	**FT Fraco LR**
Profissao	0.2065	0.2129	0.2093	0.2108
TotOutflowU3	0.1566	0.1751	0.1665	0.1548
Prazo	0.1075	0.1297	0.1318	0.1264
Idade	0.1187	0.1350	0.1351	0.1424
Variável	**CAT (Ref.)**	**FT Forte CAT**	**FT Médio CAT**	**FT Fraco CAT**
Profissao	0.2508	0.3226	0.2669	0.2817
TotOutflowU3	0.2312	0.2241	0.2135	0.2000
Prazo	0.1480	0.1647	0.1776	0.1684
Idade	0.1576	0.1740	0.1821	0.1916

O *Attribute Inference* no DP FT SD revela o problema fundamental deste método. Ao gerar cada variável de forma independente, as distribuições marginais são preservadas mas as correlações entre variáveis não o são. O problema é que o *Attribute Inference* explora precisamente essas correlações estruturais: um atacante que saiba o valor de todas as outras variáveis de um indivíduo consegue inferir o valor de uma variável sensível porque o modelo aprendeu as associações entre variáveis e a variável alvo, associações essas que estão presentes mesmo nos dados sintéticos gerados a partir das distribuições marginais reais. A Profissao mantém valores similares ao modelo de referência em todos os níveis no LR, e no *CatBoost* agrava-se significativamente no FT Forte (0.323). O TotOutflowU3 não melhora de forma consistente em nenhum dos modelos.

O DP FT SD confirmou que preservar distribuições marginais é insuficiente para reduzir a fuga de informação estrutural, pois as correlações entre variáveis e a variável alvo, que estão na origem da fuga de informação, não são capturadas por um método de amostragem independente. A questão natural que se seguiu foi se modelos generativos mais expressivos, capazes de aprender e reproduzir essas correlações de forma privatizada, conseguiriam resultados diferentes. A secção seguinte avalia métodos de geração condicional com Privacidade Diferencial, que abordam precisamente este problema através de redes neuronais generativas.

4.3.2 Geração Condicional e Privacidade Diferencial

O DP FT SD demonstrou que preservar distribuições marginais é insuficiente para capturar as correlações estruturais entre variáveis e variável alvo que estão na origem da fuga de informação. A questão que se seguiu foi se modelos generativos mais expressivos, capazes de aprender e reproduzir essas correlações, conseguiriam resultados diferentes quando combinados com garantias formais de Privacidade Diferencial.

Esta secção descreve a progressão de três abordagens nessa direção: o DP-CTGAN, que integra Privacidade Diferencial diretamente no treino de uma rede neuronal generativa condicional para dados tabulares; o CTGAN, que avalia a qualidade da geração condicional sem restrições de privacidade; e a combinação CTGAN+DPGAN, que separa explicitamente o problema do desbalanceamento do problema da privacidade.

4.3.2.1 DP-CTGAN

As *Generative Adversarial Networks* (GANs) são modelos de aprendizagem profunda compostos por dois componentes que competem entre si: um gerador, que aprende a produzir dados artificiais, e um discriminador, que aprende a distinguir dados reais de dados sintéticos. Através desta competição adversarial, o gerador é progressivamente forçado a gerar dados cada vez mais realistas. O CTGAN é uma variante das GANs desenvolvida especificamente para dados tabulares, introduzindo um mecanismo de amostragem condicional que permite controlar a geração por categoria de uma variável, incluindo a variável alvo [54].

O DP-CTGAN incorpora Privacidade Diferencial no processo de treino do CTGAN, adicionando ruído calibrado aos gradientes do discriminador através do mecanismo *DP-SGD* [36]. Embora na literatura existam implementações de DP-CTGAN que mantêm o mecanismo de geração condicional [83], as implementações disponíveis em bibliotecas Python de referência para síntese de dados com Privacidade Diferencial não incluem este mecanismo, uma vez que a integração com *DP-SGD* aumenta significativamente a complexidade da arquitetura, e as implementações disponíveis optaram por remover o vetor condicional para simplificar essa integração.

Sem geração condicional, o *DP-SGD* apresenta uma incompatibilidade estrutural com conjuntos de dados com desbalanceamento acentuado. O ruído adicionado aos gradientes de forma uniforme suprime o sinal das regiões de baixa densidade de forma desproporcional, precisamente onde se encontram os exemplos de incumprimento. O sinal da classe minoritária, já fraco por natureza num conjunto de dados com 4.4% de incumprimentos, é progressivamente suprimido ao ponto de o gerador convergir para a distribuição da classe maioritária, gerando quase exclusivamente não-incumprimentos.

Este comportamento foi confirmado empiricamente. Após o treino do DP-CTGAN, foram geradas 1 000 000 amostras sintéticas para avaliar a distribuição da variável alvo. A taxa de incumprimento no conjunto real é de 4.39%, enquanto nos dados sintéticos gerados é de apenas 0.014%, o que significa que seriam necessárias aproximadamente 12.7 milhões de amostras sintéticas para

obter um número equivalente de incumprimentos ao conjunto de treino original, tornando a abordagem computacionalmente impraticável. O DP-CTGAN foi descartado sem avaliação formal de utilidade ou privacidade. Esta constatação motivou uma abordagem alternativa que separasse explicitamente o problema do desbalanceamento do problema da privacidade, como descrito nas subsecções seguintes.

4.3.2.2 CTGAN

O *CTGAN* [54] foi avaliado como segunda abordagem desta família com dois objetivos distintos. O primeiro é validar que a geração condicional por classe produz dados sintéticos com qualidade suficiente para treino de modelos de avaliação de risco de crédito, servindo de base para a fase seguinte com DPGAN. O segundo é avaliar se a geração de dados sintéticos, mesmo sem garantias formais de Privacidade Diferencial, introduz alguma melhoria de privacidade como efeito colateral, uma vez que os dados sintéticos não contêm registos reais de clientes e o modelo nunca é exposto aos dados originais durante o treino.

O *CTGAN* foi treinado sobre os dados reais, aprendendo a distribuição conjunta das variáveis e da variável alvo. O mecanismo de amostragem condicional permite controlar explicitamente a proporção de incumprimentos nos dados sintéticos gerados, o que é fundamental para lidar com o desbalanceamento acentuado do conjunto de dados. Após o treino, foram gerados três conjuntos de dados sintéticos com diferentes proporções da classe minoritária, de forma a avaliar o impacto do nível de balanceamento na utilidade do modelo:

- **CTGAN Original** — proporção original dos dados reais (4.4% de incumprimentos)
- **CTGAN Intermédio** — proporção intermédia ($\sim$10% de incumprimentos)
- **CTGAN Balanceado** — conjunto de dados completamente balanceado (50% de incumprimentos)

O modelo de avaliação de risco de crédito é treinado exclusivamente sobre cada um destes conjuntos de dados sintéticos e avaliado sobre os dados de teste reais, seguindo a mesma metodologia das restantes experiências.

A Tabela 4.11 apresenta as métricas de utilidade e privacidade global.

Tabela 4.11: Métricas de utilidade e privacidade global — CTGAN

Modelo	AUC-ROC	KS	G-Mean	MIA	Canary Gap
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
CTGAN Original LR	0.7628	0.4217	0.7013	0.6051	1.26
CTGAN Intermédio LR	0.7675	0.4248	0.7098	0.6135	1.09
CTGAN Balanceado LR	0.7636	0.4259	0.7096	0.6407	0.59
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
CTGAN Original CAT	0.7550	0.4112	0.6936	0.6189	0.64
CTGAN Intermédio CAT	0.7582	0.4094	0.6955	0.6301	0.67
CTGAN Balanceado CAT	0.7589	0.4204	0.7045	0.6567	0.44

Os resultados confirmam que o *CTGAN* gera dados sintéticos com utilidade próxima do modelo de referência. O CTGAN Intermédio LR tem AUC-ROC de 0.768, apenas 0.011 abaixo do modelo de referência, e G-Mean de 0.710, praticamente igual. O *CatBoost* segue o mesmo padrão, com AUC-ROC entre 0.755 e 0.759. Este resultado valida que a geração condicional produz dados sintéticos com qualidade suficiente para treino de modelos de avaliação de risco de crédito, independentemente do nível de balanceamento escolhido.

O Canary Gap desce em todas as configurações face ao modelo de referência, situando-se entre 0.59 e 1.26 no LR e entre 0.44 e 0.67 no *CatBoost*. Tal como observado no DP FT SD, um modelo treinado sobre dados sintéticos aprende distribuições artificiais que reduzem a capacidade de memorizar combinações de valores específicas presentes nos dados reais.

O MIA sobe para 0.605–0.641, refletindo o desfasamento entre dados sintéticos de treino e dados reais de teste, o mesmo fenómeno observado em todos os métodos anteriores.

O *Attribute Inference* mantém-se próximo do modelo de referência em todas as configurações, com Profissao entre 0.203 e 0.291 e TotOutflowU3 entre 0.144 e 0.178. Este resultado confirma que a geração de dados sintéticos, mesmo quando condicional e de alta qualidade, não introduz melhoria de privacidade como efeito colateral, pois as correlações estruturais entre variáveis e variável alvo que estão na origem da fuga de informação são precisamente o que o *CTGAN* aprende e reproduz. Os dados sintéticos têm boa utilidade exatamente porque preservam essas correlações, e é por essa razão que não reduzem o *Attribute Inference*.

Esta constatação coloca a questão de se adicionar garantias formais de Privacidade Diferencial ao processo de geração conseguiria quebrar essas correlações estruturais sem destruir a utilidade do modelo. A DPGAN, avaliada na subsecção seguinte, testa precisamente esta hipótese.

4.3.2.3 DPGAN

A DPGAN [55] é uma variante das GANs que incorpora Privacidade Diferencial no processo de treino, adicionando ruído calibrado aos gradientes do discriminador através do mecanismo *DP-SGD*. Tal como o CTGAN, a DPGAN gera dados sintéticos, mas com garantias formais de privacidade. O modelo de avaliação de risco de crédito é treinado exclusivamente sobre os dados sintéticos gerados pela DPGAN, sem acesso aos dados reais.

Esta abordagem funciona em duas fases, ilustradas na Figura 4.1. Numa primeira fase, o conjunto de dados de treino original é balanceado com dados sintéticos gerados pelo CTGAN, adicionando exclusivamente exemplos da classe minoritária até atingir o nível de balanceamento pretendido. Numa segunda fase, a DPGAN é treinada sobre este conjunto de dados balanceado e gera um novo conjunto de dados sintético com garantias formais de Privacidade Diferencial, sobre o qual o modelo de avaliação de risco de crédito é finalmente treinado.

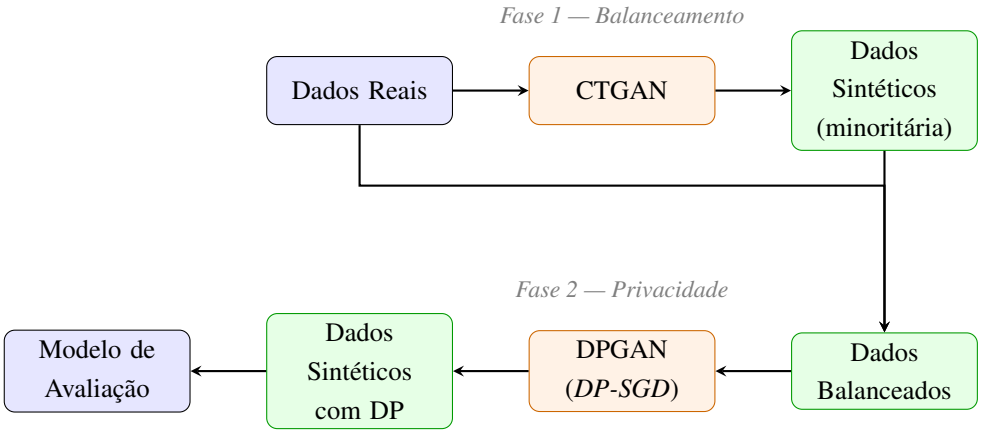

Figura 4.1: Fluxo de processamento da abordagem CTGAN+DPGAN

Foram testadas três proporções de balanceamento (10%, 30% e 50% de incumprimentos), de forma a cobrir o espectro entre um balanceamento mínimo, que preserva a distribuição original, e um balanceamento total. Para o parâmetro de privacidade, foram escolhidos dois níveis ($\varepsilon = 2$ e $\varepsilon = 5$), correspondendo a níveis de proteção moderados, dentro do intervalo tipicamente explorado na literatura de síntese de dados com Privacidade Diferencial [55]. Esta configuração inicial pretendia avaliar se a abordagem era viável antes de uma exploração mais detalhada do espaço de parâmetros, exploração que se revelou desnecessária face à clareza dos resultados obtidos.

A Tabela 4.12 apresenta as métricas de utilidade e privacidade global para todas as configurações.

Tabela 4.12: Métricas de utilidade e privacidade global — DPGAN

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
DPGAN r01 ε=2 LR	0.6127	0.1752	0.2638	0.9415	0.87
DPGAN r01 ε=5 LR	0.6508	0.2352	0.2265	0.9285	0.91
DPGAN r03 ε=2 LR	0.6357	0.2058	0.5579	0.9548	0.73
DPGAN r03 ε=5 LR	0.5377	0.0969	0.5287	0.8744	1.45
DPGAN r05 ε=2 LR	0.5283	0.0629	0.5124	0.9488	0.65
DPGAN r05 ε=5 LR	0.5443	0.1218	0.5304	0.9553	0.83
DPGAN r01 ε=2 CAT	0.5874	0.1603	0.5050	0.9588	0.62
DPGAN r01 ε=5 CAT	0.6682	0.2497	0.3617	0.9486	0.83
DPGAN r03 ε=2 CAT	0.6476	0.2285	0.5530	0.9620	0.72
DPGAN r03 ε=5 CAT	0.5423	0.0914	0.5224	0.9124	0.84
DPGAN r05 ε=2 CAT	0.5244	0.0505	0.5179	0.9572	0.39
DPGAN r05 ε=5 CAT	0.5209	0.0466	0.4744	0.9697	0.75

Os resultados são inequívocos: os dados sintéticos gerados pela DPGAN destroem a utilidade do modelo de avaliação de risco de crédito em todas as configurações testadas. O melhor resultado obtido foi o DPGAN r01 ε=5 LR com AUC-ROC de 0.651, muito abaixo do aceitável para uso em produção. O G-Mean extremamente baixo em várias configurações confirma que os modelos

treinados sobre dados sintéticos DPGAN têm dificuldade em detetar incumprimentos de forma equilibrada entre as duas classes.

O MIA atinge valores entre 0.874 e 0.970, os valores mais altos de todos os métodos testados, refletindo o desfasamento extremo entre os dados sintéticos gerados pela DPGAN e os dados de teste reais, amplificado pelo ruído de Privacidade Diferencial adicionado aos gradientes durante a geração.

O mecanismo subjacente é claro: o ruído de Privacidade Diferencial adicionado aos gradientes durante o treino da DPGAN degrada as correlações entre variáveis e variável alvo a um ponto em que o modelo de avaliação de risco de crédito não consegue aprender os padrões necessários. Este efeito agrava-se com o aumento do nível de balanceamento, uma vez que mais exemplos da classe minoritária amplificam o desfasamento entre os dados sintéticos gerados e os dados de teste reais.

Face à incapacidade da DPGAN em gerar dados sintéticos com utilidade aceitável, a investigação abandonou a família de dados sintéticos com Privacidade Diferencial. Os resultados desta família revelaram um padrão consistente: a Privacidade Diferencial aplicada ao processo de geração de dados degrada invariavelmente as correlações estruturais entre variáveis e variável alvo, que são precisamente as que os modelos de avaliação de risco de crédito precisam de aprender. Uma abordagem alternativa é não intervir nos dados, mas sim proteger o próprio processo de treino do modelo, garantindo que o modelo não memoriza exemplos individuais durante o treino, independentemente dos dados sobre os quais é treinado. Esta família de métodos é avaliada na secção seguinte.

4.4 Proteção do Modelo

As abordagens de proteção do modelo inserem-se na família de técnicas de PPML ao nível do modelo e do treino, descrita na Secção 2.2.2. Em vez de atuar sobre os dados antes do treino, estas técnicas modificam o próprio processo de aprendizagem para limitar a informação que o modelo retém sobre os exemplos individuais de treino. A motivação para explorar esta família surge dos resultados obtidos com os dados sintéticos, onde ficou evidente que atuar ao nível dos dados não resolve a fuga de informação estrutural, levantando a questão de se proteger diretamente o modelo durante o treino poderia produzir resultados diferentes.

Nesta secção são avaliados três métodos: Privacidade Diferencial no treino do classificador, PATE e Fairlearn.

4.4.1 Privacidade Diferencial no Treino do Classificador

A Privacidade Diferencial pode ser incorporada diretamente no processo de treino do classificador através do mecanismo *DP-SGD* [36]. Em cada iteração do treino, os gradientes calculados para cada exemplo individual são truncados para um valor máximo, limitando a influência de qualquer exemplo individual, e é adicionado ruído gaussiano calibrado pelo parâmetro ε antes da atualização dos parâmetros do modelo, com garantia formal ε-DP.

Esta abordagem revelou-se, contudo, incompatível com dados de codificação binária multivariada neste contexto: a avaliação empírica mostrou que a degradação de utilidade é severa e independente do valor de epsilon, confirmando que o mecanismo de truncagem e ruído não se adapta adequadamente a esta representação dos dados. Dado que este contexto combina essa codificação com discretização e desbalanceamento acentuado, foi conduzida uma avaliação empírica para confirmar o comportamento neste cenário específico. A alternativa disponível compatível com dados tabulares categóricos foi uma implementação de *Random Forest* com Privacidade Diferencial.

Foram testados quatro níveis de privacidade ($\varepsilon \in \{1.0, 5.0, 10.0, 20.0\}$), de forma a avaliar o impacto do epsilon na utilidade e privacidade do modelo. A Tabela 4.13 apresenta os resultados obtidos.

Tabela 4.13: Métricas de utilidade e privacidade global — DP *Random Forest*

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
DP RF ε=1.0	0.5982	0.1558	0.0000	0.4961	0.28
DP RF ε=5.0	0.5870	0.1479	0.0000	0.4978	-0.07
DP RF ε=10.0	0.5862	0.1462	0.0000	0.4970	-0.07
DP RF ε=20.0	0.5862	0.1462	0.0000	0.4970	-0.07

Os resultados confirmam empiricamente as limitações documentadas na literatura. O AUC-ROC situa-se entre 0.586 e 0.598 e o G-Mean é nulo em todas as configurações, indicando que o modelo não deteta nenhum incumprimento. As métricas de privacidade e utilidade são praticamente idênticas para todos os níveis de epsilon, confirmando que o mecanismo é insensível à variação do parâmetro de privacidade neste contexto, precisamente pelo motivo documentado: a uniformidade dos limites $[0, 1]$ em dados com codificação binária multivariada torna a sensibilidade constante independentemente do valor de ε.

O PATE, avaliado na secção seguinte, aborda este problema de forma diferente: em vez de adicionar ruído aos gradientes do modelo, agrega as previsões de múltiplos modelos professores com garantias formais de privacidade, sem depender da natureza contínua ou categórica dos dados.

4.4.2 PATE — *Private Aggregation of Teacher Ensembles*

O *Private Aggregation of Teacher Ensembles* (PATE) [59] é uma abordagem de PPML ao nível do modelo que incorpora Privacidade Diferencial no processo de transferência de conhecimento entre modelos. O mecanismo funciona em duas fases: primeiro, treina-se um conjunto de modelos professores sobre partições disjuntas dos dados de treino, em que cada professor vê apenas uma parte dos dados, limitando a informação que cada um retém sobre exemplos individuais. Depois, as previsões dos professores são agregadas com adição de ruído para treinar um modelo aluno, que aprende a imitar o comportamento coletivo dos professores sem nunca ter acesso direto aos dados reais. A garantia de privacidade resulta precisamente desta separação: o aluno aprende apenas a partir das previsões ruidosas dos professores.

A motivação para explorar o PATE surge da sua abordagem concetualmente diferente face ao *DP-SGD*: em vez de adicionar ruído diretamente aos gradientes durante o treino, o PATE limita a informação disponível ao modelo aluno através do particionamento dos dados e da agregação ruidosa. Esta diferença é relevante porque os resultados anteriores mostraram que o ruído nos gradientes destrói a capacidade de aprendizagem do modelo.

Foram testadas 20 configurações combinando cinco valores para o número de professores ($T \in \{3,4,5,6,7\}$) e quatro escalas de ruído na agregação ($N \in \{0.1,0.2,0.3,0.5\}$). O intervalo de valores para o número de professores foi escolhido de forma a equilibrar dois efeitos opostos documentados na literatura: um número maior de professores aumenta a robustez das previsões agregadas, tornando a votação mais consistente, mas reduz simultaneamente os dados disponíveis para treinar cada professor individualmente, o que pode comprometer a qualidade das previsões de cada modelo [59]. O intervalo de escalas de ruído cobre o espectro entre perturbação baixa, que preserva mais utilidade, e perturbação elevada, que oferece maiores garantias de privacidade.

As Tabelas 4.14 e 4.15 apresentam, respetivamente, as métricas de utilidade e privacidade global, e o *Attribute Inference* nas variáveis de maior fuga de informação, para uma seleção representativa das 20 configurações testadas.

Tabela 4.14: Métricas de utilidade e privacidade global — PATE (seleção representativa)

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
PATE T3 N0.1	0.7663	0.4155	0.7052	0.5026	5.53
PATE T3 N0.3	0.7671	0.4147	0.7038	0.5045	5.97
PATE T3 N0.5	0.7665	0.4150	0.7034	0.5046	5.72
PATE T5 N0.1	0.7673	0.4116	0.7026	0.4990	6.27
PATE T5 N0.3	0.7678	0.4138	0.7033	0.5024	5.17
PATE T7 N0.1	0.7693	0.4225	0.7081	0.5031	5.67
PATE T7 N0.3	0.7691	0.4213	0.7084	0.5019	4.51

Tabela 4.15: *Attribute Inference* — PATE (seleção representativa)

Modelo	**Profissao**	**TotOutflowU3**	**Prazo**	**Idade**
CAT (Ref.)	0.251	0.231	0.148	0.158
PATE T3 N0.1	0.329	0.251	0.155	0.116
PATE T3 N0.3	0.316	0.289	0.158	0.084
PATE T3 N0.5	0.324	0.243	0.146	0.102
PATE T5 N0.1	0.313	0.250	0.165	0.080
PATE T5 N0.3	0.313	0.245	0.162	0.111
PATE T7 N0.1	0.318	0.240	0.181	0.117
PATE T7 N0.3	0.312	0.239	0.174	0.119

Os resultados do PATE confirmam que o método resolve eficazmente o problema para que foi desenhado: a AUC-ROC mantém-se entre 0.766 e 0.769 em todas as configurações, com degradação mínima face ao modelo de referência, o MIA mantém-se próximo de 0.50 e o Canary Gap é consistentemente elevado em todas as configurações testadas, entre 4.06 e 6.27, confirmando que o PATE não memoriza exemplos individuais.

No entanto, estes resultados positivos revelam precisamente a limitação do PATE neste contexto. A Tabela 4.15 mostra que o *Attribute Inference* nas variáveis de maior fuga de informação não melhora face ao modelo de referência: a Profissao sobe de 0.251 para valores entre 0.312 e 0.330, o TotOutflowU3 de 0.231 para valores entre 0.227 e 0.289, o Prazo de 0.148 para valores entre 0.146 e 0.181, e a Idade de 0.158 para valores entre 0.080 e 0.166. Esta aparente contradição, ou seja, um método com fortes garantias de privacidade contra memorização que simultaneamente não reduz a fuga de informação por inferência de atributos, tem uma explicação estrutural. O PATE foi concebido para proteger exemplos individuais: o particionamento dos dados e a agregação ruidosa garantem que nenhum registo específico influencia demasiado o modelo aluno. Mas a fuga de informação por inferência de atributos não resulta da memorização de exemplos individuais, mas sim das correlações populacionais entre variáveis como a Profissao e o TotOutflowU3 e o risco de incumprimento, que existem nos dados independentemente de como o modelo foi treinado. Qualquer modelo com boa capacidade preditiva aprende essas correlações, e o PATE não é exceção. Proteger o processo de treino não resolve um problema que tem origem na estrutura dos próprios dados.

Esta constatação motivou a exploração de uma abordagem diferente: em vez de proteger o processo de treino, aplicar restrições explícitas de equidade sobre as previsões do modelo. A aprendizagem com restrições de equidade, avaliada na secção seguinte, testa esta hipótese.

4.4.3 Aprendizagem com Restrições de Equidade

A aprendizagem com restrições de equidade permite treinar modelos com restrições explícitas sobre variáveis sensíveis, forçando o modelo a tratar diferentes grupos de forma mais equitativa nas suas previsões. A motivação para explorar esta abordagem surgiu da análise dos resultados do PATE: se os métodos de proteção do modelo não resolvem a fuga de informação estrutural porque atacam um problema diferente, a hipótese era que forçar o modelo a tratar os grupos de uma variável sensível de forma mais equitativa poderia reduzir as diferenças entre grupos e, consequentemente, reduzir a fuga de informação associada a essa variável.

É importante sublinhar que esta foi uma hipótese exploratória, pois equidade e privacidade são conceitos distintos na literatura, e não existe garantia teórica de que restrições de equidade impliquem redução de fuga de informação [62]. No entanto, a intuição era plausível e justificou a experimentação.

Foram testadas oito configurações, combinando quatro variáveis sensíveis (Profissao, TotOutflowU3, Prazo e Idade) com duas restrições de equidade: *DemographicParity*, que exige que as taxas de decisão positiva sejam iguais entre grupos, e *EqualizedOdds*, que exige que as taxas de verdadeiros positivos e falsos positivos sejam iguais entre grupos. As variáveis foram escolhidas

por serem as quatro com maior fuga de informação no modelo de referência, sendo precisamente sobre estas que se pretende avaliar o impacto das restrições de equidade.

As Tabelas 4.16 e 4.17 apresentam, respetivamente, as métricas de utilidade e privacidade global, e o *Attribute Inference* nas variáveis de maior fuga de informação.

Tabela 4.16: Métricas de utilidade e privacidade global — Aprendizagem com Restrições de Equidade

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
CAT (Ref.)	0.781	0.436	0.712	0.503	1.90
FL Profissao DP	0.677	0.341	0.538	0.499	1.39
FL Profissao EO	0.719	0.389	0.447	0.499	0.38
FL TotOutflowU3 DP	0.739	0.443	0.695	0.504	7.67
FL TotOutflowU3 EO	0.727	0.433	0.538	0.501	0.06
FL Prazo DP	0.705	0.430	0.643	0.500	1.05
FL Prazo EO	0.727	0.428	0.641	0.499	1.07
FL Idade DP	0.735	0.393	0.674	0.494	18.90
FL Idade EO	0.735	0.388	0.522	0.496	0.10

Tabela 4.17: *Attribute Inference* — Aprendizagem com Restrições de Equidade

Modelo	**Profissao**	**TotOutflowU3**	**Prazo**	**Idade**
CAT (Ref.)	0.251	0.231	0.148	0.158
FL Profissao DP	0.447	0.048	0.022	0.033
FL Profissao EO	0.494	0.062	0.055	0.049
FL TotOutflowU3 DP	0.051	0.242	0.020	0.018
FL TotOutflowU3 EO	0.093	0.109	0.031	0.016
FL Prazo DP	0.030	0.014	0.228	0.007
FL Prazo EO	0.073	0.039	0.234	0.026
FL Idade DP	0.099	0.026	0.016	0.089
FL Idade EO	0.126	0.050	0.029	0.399

Os resultados revelam um padrão paradoxal e consistente em todas as configurações testadas: ao aplicar restrições de equidade sobre uma variável sensível, a fuga de informação dessa variável aumenta em vez de diminuir. Quando se aplica *DemographicParity* sobre a Profissao, o *Attribute Inference* sobe de 0.251 para 0.447, quase o dobro do valor de referência. O mesmo padrão repete-se nas restantes variáveis: o TotOutflowU3 mantém-se em 0.242 com *DemographicParity* e sobe para 0.109 com *EqualizedOdds*, o Prazo sobe para 0.228 e 0.234, e a Idade sobe para 0.089 e 0.399 respetivamente. Em todos os casos, as outras variáveis, aquelas sobre as quais não são impostas restrições, registam uma redução acentuada da fuga de informação, o que confirma que o efeito é específico da variável restringida e não um artefacto do método.

O mecanismo subjacente explica o paradoxo: ao forçar o modelo a tratar os grupos de forma mais equitativa, o modelo torna-se internamente mais sensível às diferenças entre grupos para

conseguir cumprir a restrição. O modelo aprende mais sobre a variável sensível, e o atacante consegue explorar exatamente essa informação adicional.

Este resultado confirma empiricamente uma distinção documentada na literatura: equidade e privacidade são objetivos distintos e potencialmente contraditórios, e cumprir restrições de equidade não implica proteger a privacidade das variáveis sensíveis [62].

Os resultados desta família de métodos (Privacidade Diferencial no treino, PATE e aprendizagem com restrições de equidade) apontam para uma conclusão consistente: as técnicas que atuam ao nível do modelo reduzem a memorização de exemplos individuais mas não conseguem eliminar a fuga de informação que tem origem nas correlações estruturais entre variáveis e variável alvo. Esta conclusão motivou um regresso à perturbação direta dos dados, desta vez com uma abordagem diferente da explorada anteriormente, descrita na secção seguinte.

4.5 Perturbação por Ruído Simples

Os métodos avaliados nas secções anteriores (mecanismos com garantia formal de Privacidade Diferencial, geração de dados sintéticos e proteção do modelo) revelaram um padrão comum neste contexto: a complexidade dos mecanismos e a força das restrições de privacidade acabam por trabalhar contra as características específicas deste conjunto de dados. Dados reais de avaliação de risco de crédito, com correlações fortes entre variáveis e variável alvo e um desbalanceamento acentuado entre classes, são particularmente sensíveis a perturbações agressivas: o sinal preditivo da classe minoritária é suprimido, as correlações estruturais são destruídas, e o modelo perde a capacidade de aprender os padrões necessários. Os métodos de dados sintéticos, por sua vez, preservam essas correlações quando funcionam bem, mas é precisamente por isso que não reduzem a fuga de informação.

Esta observação motivou uma abordagem diferente: em vez de mecanismos sofisticados com múltiplos parâmetros e garantias formais complexas, testar perturbações simples e diretas sobre os dados. A intuição é que perturbações mais simples introduzem ruído mais previsível e controlável, sem os efeitos secundários que os métodos anteriores revelaram neste contexto específico.

Nesta secção são avaliados três métodos: Perturbação Mista Pré-Discretização, que aplica mecanismos de ruído distintos por tipo de variável antes da discretização; Substituição Aleatória de Categorias (*Random Category Substitution*, RCS) uniforme; e Troca de Categorias (*Category Swap*) uniforme. Cada método é avaliado com diferentes intensidades de perturbação, de forma a caracterizar o compromisso entre privacidade e utilidade ao longo de um espectro de ruído controlado.

4.5.1 Perturbação Mista Pré-Discretização

Este método aplica perturbação diretamente sobre os dados originais antes da discretização, utilizando mecanismos distintos adaptados ao tipo de cada variável. As variáveis numéricas são perturbadas com ruído Gaussiano ou de Laplace, adicionando a cada valor uma perturbação amostrada de uma distribuição centrada em zero com escala proporcional ao desvio padrão da variável.

As variáveis categóricas são perturbadas através de Substituição Aleatória de Categorias (*Random Category Substitution*), substituindo o valor original por uma categoria alternativa escolhida aleatoriamente. As variáveis binárias são perturbadas através de inversão de bit (*Bit Flipping*), invertendo o valor com uma determinada probabilidade.

A motivação para aplicar a perturbação antes da discretização é alargar o conjunto de mecanismos disponíveis: as variáveis numéricas contínuas permitem a aplicação de ruído contínuo, que não seria aplicável após a discretização. No entanto, como foi discutido na Secção 3.2, esta escolha expõe a perturbação ao risco de ser absorvida ou distorcida pelo processo de discretização subsequente.

Foram testados quatro níveis de intensidade ($\sigma \in \{0.05, 0.10, 0.20, 0.30\}$), aplicados em simultâneo a todos os tipos de variáveis: nas numéricas como escala do ruído, nas categóricas como probabilidade de substituição, e nas binárias como probabilidade de inversão (com metade do valor de σ), totalizando oito configurações.

A Tabela 4.18 apresenta as métricas de utilidade e privacidade global para uma seleção representativa de configurações.

Tabela 4.18: Métricas de utilidade e privacidade global — Perturbação Mista Pré-Discretização

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
Gaussiano σ=0.05 LR	0.7695	0.4208	0.7062	0.6534	16.38
Gaussiano σ=0.10 LR	0.7669	0.4186	0.7014	0.6556	14.46
Gaussiano σ=0.20 LR	0.7688	0.4286	0.7078	0.6576	12.85
Gaussiano σ=0.30 LR	0.7704	0.4234	0.7063	0.6648	10.86
Laplace σ=0.05 LR	0.7760	0.4400	0.7152	0.6506	15.71
Laplace σ=0.10 LR	0.7753	0.4329	0.7127	0.6559	15.09
Laplace σ=0.20 LR	0.7651	0.4221	0.7023	0.6602	13.81
Laplace σ=0.30 LR	0.7621	0.4175	0.6811	0.6731	11.45
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
Gaussiano σ=0.05 CAT	0.7739	0.4351	0.7126	0.6623	1.34
Gaussiano σ=0.10 CAT	0.7685	0.4213	0.7051	0.6611	0.13
Gaussiano σ=0.20 CAT	0.7645	0.4307	0.7111	0.6516	0.94
Gaussiano σ=0.30 CAT	0.7572	0.3948	0.6833	0.6743	0.41
Laplace σ=0.05 CAT	0.7718	0.4250	0.7097	0.6585	3.23
Laplace σ=0.10 CAT	0.7704	0.4281	0.7088	0.6579	1.28
Laplace σ=0.20 CAT	0.7433	0.3777	0.6644	0.6845	1.18
Laplace σ=0.30 CAT	0.7619	0.4073	0.6826	0.6771	0.44

A degradação de utilidade é moderada em ambos os modelos: mesmo com $\sigma = 0.30$, o LR mantém AUC-ROC de 0.770 e o *CatBoost* de 0.757. O Canary Gap do LR mantém-se elevado em todas as configurações, entre 10.9 e 16.4, indicando ausência de memorização. No *CatBoost*, o Canary Gap é mais variável e desce para valores próximos de zero em algumas configurações

Gaussianas, sugerindo alguma tendência de memorização. O MIA sobe para 0.65–0.67 em ambos os modelos, refletindo o desfasamento entre dados perturbados de treino e dados reais de teste.

A Tabela 4.19 apresenta os resultados do *Attribute Inference* para as variáveis de maior fuga de informação.

Tabela 4.19: *Attribute Inference* — Perturbação Mista Pré-Discretização (variáveis de maior fuga de informação)

Modelo	**Profissao**	**TotOutflowU3**	**Prazo**	**Idade**
LR (Ref.)	0.2065	0.1566	0.1075	0.1187
Gauss. σ=0.05 LR	0.2393	0.1843	0.1056	0.1407
Gauss. σ=0.10 LR	0.2391	0.1750	0.1006	0.1364
Gauss. σ=0.20 LR	0.2350	0.1712	0.1250	0.1479
Gauss. σ=0.30 LR	0.2333	0.1912	0.1335	0.1462
Lap. σ=0.05 LR	0.2400	0.1865	0.0946	0.1379
Lap. σ=0.10 LR	0.2677	0.1938	0.0996	0.1422
Lap. σ=0.20 LR	0.2471	0.1935	0.1153	0.1378
Lap. σ=0.30 LR	0.2135	0.1699	0.1170	0.1359
CAT (Ref.)	0.2508	0.2312	0.1480	0.1576
Gauss. σ=0.05 CAT	0.3039	0.2877	0.1037	0.2060
Gauss. σ=0.10 CAT	0.2860	0.2404	0.0938	0.1516
Gauss. σ=0.20 CAT	0.3082	0.2583	0.1007	0.1778
Gauss. σ=0.30 CAT	0.2998	0.2269	0.0832	0.1679
Lap. σ=0.05 CAT	0.3021	0.2862	0.0921	0.1551
Lap. σ=0.10 CAT	0.3125	0.2688	0.0733	0.1525
Lap. σ=0.20 CAT	0.3172	0.2839	0.0634	0.1598
Lap. σ=0.30 CAT	0.2899	0.2318	0.1037	0.1579

O *Attribute Inference* revela o problema fundamental deste método. Ao perturbar os valores antes da discretização, o ruído é parcialmente absorvido pelo processo de discretização subsequente, pelo que as correlações estruturais entre variáveis e variável alvo mantêm-se praticamente intactas após a discretização, e a fuga de informação não melhora. No LR, a Profissao e o TotOutflowU3 sobem em relação ao modelo de referência em todas as configurações. No *CatBoost* o padrão é ainda mais acentuado: a Profissao sobe para valores entre 0.290 e 0.317, e o TotOutflowU3 mantém-se acima de 0.226 em todas as configurações.

Este resultado confirma empiricamente a limitação identificada na Secção 3.2: a perturbação aplicada antes da discretização é absorvida ou distorcida pela discretização, tornando este ponto de aplicação pouco eficaz para reduzir a fuga de informação por inferência de atributos. Face a este resultado, a investigação orientou-se para perturbação direta sobre as variáveis já discretizadas, explorada nas secções seguintes.

4.5.2 Substituição Aleatória de Categorias

A Substituição Aleatória de Categorias (*Random Category Substitution*, RCS) [48] é um método de perturbação categórica que, para cada registo, substitui o valor de uma variável por uma categoria diferente escolhida aleatoriamente de entre as categorias possíveis dessa variável, com probabilidade p. O parâmetro $p \in [0,1]$ controla a intensidade da perturbação: para $p = 0$ os dados mantêm-se inalterados, e para $p = 1$ todos os valores são substituídos por uma categoria aleatória diferente. A perturbação é aplicada independentemente a cada variável e a cada registo, após a discretização.

Ao contrário dos mecanismos LDP como o GRR, o RCS não oferece garantias formais de privacidade, sendo uma perturbação heurística sem calibração teórica. No entanto, a sua simplicidade permite um controlo direto e intuitivo sobre a intensidade da perturbação através do parâmetro p.

Foram testados cinco níveis de perturbação ($p \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$), aplicados às 14 variáveis discretizadas.

A Tabela 4.20 apresenta as métricas de utilidade e privacidade global.

Tabela 4.20: Métricas de utilidade e privacidade global — RCS

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
RCS p=0.05 LR	0.7796	0.4419	0.7153	0.5301	13.68
RCS p=0.10 LR	0.7762	0.4413	0.7186	0.5716	11.24
RCS p=0.20 LR	0.7752	0.4456	0.7173	0.6288	9.51
RCS p=0.30 LR	0.7648	0.4285	0.7019	0.6701	6.01
RCS p=0.50 LR	0.7481	0.3840	0.6762	0.7150	2.95
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
RCS p=0.05 CAT	0.7787	0.4459	0.7166	0.5351	1.08
RCS p=0.10 CAT	0.7733	0.4384	0.7162	0.5633	0.84
RCS p=0.20 CAT	0.7691	0.4261	0.7077	0.6409	1.08
RCS p=0.30 CAT	0.7580	0.4101	0.6962	0.6674	0.37
RCS p=0.50 CAT	0.7330	0.3727	0.6605	0.7005	0.12

O RCS mantém boa utilidade para valores baixos de p: o RCS p=0.05 LR tem AUC-ROC de 0.780, praticamente igual ao modelo de referência, e o *CatBoost* segue o mesmo padrão com AUC-ROC de 0.779. À medida que p aumenta, a utilidade degrada progressivamente em ambos os modelos, mas de forma mais suave que os métodos LDP. O MIA sobe com p, refletindo o mesmo desfasamento de distribuição observado nos métodos anteriores.

A Tabela 4.21 apresenta os resultados do *Attribute Inference* para as variáveis de maior fuga de informação.

Tabela 4.21: *Attribute Inference* — RCS (variáveis de maior fuga de informação)

Modelo	**Profissao**	**TotOutflowU3**	**Prazo**	**Idade**
LR (Ref.)	0.207	0.157	0.108	0.119
RCS p=0.05 LR	0.214	0.153	0.114	0.118
RCS p=0.10 LR	0.223	0.175	0.077	0.125
RCS p=0.20 LR	0.215	0.156	0.086	0.131
RCS p=0.30 LR	0.271	0.188	0.099	0.165
RCS p=0.50 LR	0.241	0.185	0.089	0.176
CAT (Ref.)	0.251	0.231	0.148	0.158
RCS p=0.05 CAT	0.277	0.233	0.139	0.149
RCS p=0.10 CAT	0.256	0.237	0.103	0.156
RCS p=0.20 CAT	0.257	0.210	0.099	0.160
RCS p=0.30 CAT	0.360	0.236	0.092	0.176
RCS p=0.50 CAT	0.322	0.202	0.082	0.194

A fuga de informação por inferência de atributos não melhora com o RCS: os valores mantêm-se próximos do modelo de referência em todos os níveis de perturbação, e em alguns casos sobem significativamente. No *CatBoost*, a Profissao sobe para 0.360 com p=0.30, um agravamento considerável. A perturbação uniforme não discrimina entre variáveis com alta e baixa fuga de informação, perturbando todas igualmente, sem resolver o problema estrutural e introduzindo degradação de utilidade desnecessária nas variáveis menos problemáticas. Face a este resultado, a investigação testou uma variante diferente de perturbação categórica simples, a Troca de Categorias (*Category Swap*), que em vez de substituir valores por categorias aleatórias, troca valores entre registos, preservando a distribuição marginal de cada variável.

4.5.3 Troca de Categorias

A Troca de Categorias (*Category Swap*) [49] é um método de perturbação que troca os valores de uma variável entre pares de registos aleatórios do conjunto de dados, com probabilidade p. Ao contrário do RCS, a Troca de Categorias não introduz novos valores no conjunto de dados, apenas redistribui os valores existentes entre registos, preservando a distribuição marginal de cada variável mas quebrando as correlações entre variáveis e entre variáveis e variável alvo.

Foram testados os mesmos cinco níveis de perturbação ($p \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$), aplicados às 14 variáveis discretizadas.

A Tabela 4.22 apresenta as métricas de utilidade e privacidade global. A Troca de Categorias mantém melhor utilidade que o RCS para os mesmos níveis de perturbação: o Swap p=0.50 LR tem AUC-ROC de 0.766 face a 0.748 do RCS p=0.50 LR. Este resultado é consistente com a natureza do método, pois ao preservar a distribuição marginal de cada variável, o modelo consegue aprender melhor as estatísticas globais mesmo com perturbação elevada. O Canary Gap do LR é consistentemente mais alto na Troca de Categorias do que no RCS, indicando que este método preserva melhor a estrutura dos dados e o modelo memoriza menos os canários.

Tabela 4.22: Métricas de utilidade e privacidade global — Troca de Categorias

Modelo	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
Swap p=0.05 LR	0.7786	0.4426	0.7129	0.5213	16.90
Swap p=0.10 LR	0.7761	0.4386	0.7188	0.5451	15.18
Swap p=0.20 LR	0.7751	0.4372	0.7122	0.5901	14.12
Swap p=0.30 LR	0.7709	0.4335	0.7052	0.6198	10.17
Swap p=0.50 LR	0.7658	0.4212	0.6984	0.6350	4.68
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
Swap p=0.05 CAT	0.7802	0.4435	0.7152	0.5197	2.11
Swap p=0.10 CAT	0.7735	0.4351	0.7136	0.5475	1.48
Swap p=0.20 CAT	0.7716	0.4257	0.7117	0.6000	1.00
Swap p=0.30 CAT	0.7629	0.4008	0.6988	0.6193	1.21
Swap p=0.50 CAT	0.7485	0.3744	0.6821	0.6368	-0.06

A Tabela 4.23 apresenta os resultados do *Attribute Inference* para as variáveis de maior fuga de informação. O *Attribute Inference* na Troca de Categorias apresenta o mesmo problema do RCS: a fuga de informação não melhora de forma consistente. A Profissao reduz ligeiramente com p alto (0.160 no LR e 0.170 no *CatBoost* com p=0.50) mas à custa de degradação significativa de utilidade. O TotOutflowU3 não melhora em nenhum nível em ambos os modelos.

Tabela 4.23: *Attribute Inference* — Troca de Categorias (variáveis de maior fuga de informação)

Modelo	**Profissao**	**TotOutflowU3**	**Prazo**	**Idade**
LR (Ref.)	0.207	0.157	0.108	0.119
Swap p=0.05 LR	0.212	0.184	0.110	0.135
Swap p=0.10 LR	0.193	0.184	0.103	0.133
Swap p=0.20 LR	0.201	0.187	0.086	0.136
Swap p=0.30 LR	0.190	0.205	0.100	0.134
Swap p=0.50 LR	0.160	0.175	0.089	0.126
CAT (Ref.)	0.251	0.231	0.148	0.158
Swap p=0.05 CAT	0.258	0.241	0.153	0.151
Swap p=0.10 CAT	0.247	0.244	0.147	0.167
Swap p=0.20 CAT	0.249	0.239	0.120	0.170
Swap p=0.30 CAT	0.245	0.249	0.122	0.155
Swap p=0.50 CAT	0.170	0.221	0.092	0.147

A conclusão comum ao RCS e à Troca de Categorias é que perturbar todas as variáveis igualmente é ineficiente, pois introduz degradação de utilidade desnecessária nas variáveis menos problemáticas sem reduzir a fuga de informação nas mais críticas. Este resultado motivou a hipótese central da secção seguinte: em vez de uma perturbação uniforme, calibrar a intensidade da perturbação pela fuga de informação efetiva de cada variável, concentrando o ruído onde é de facto necessário.

4.6 Perturbação Seletiva

Os resultados obtidos ao longo das secções anteriores convergem para uma conclusão comum: a perturbação uniforme (seja por mecanismos de Privacidade Diferencial Local, por dados sintéticos, por proteção do modelo ou por perturbação simples) não resolve a fuga de informação estrutural sem sacrificar utilidade. O problema fundamental é que todas as variáveis são perturbadas com a mesma intensidade, independentemente do seu contributo para a fuga de informação. Este compromisso revela-se irresolvível na perturbação uniforme: uma intensidade baixa preserva a utilidade do modelo mas não traz qualquer benefício em termos de privacidade nas variáveis mais problemáticas; uma intensidade alta reduz a fuga de informação nessas variáveis mas destrói desnecessariamente a utilidade nas variáveis menos problemáticas, que não precisavam de ser perturbadas.

A análise do modelo de referência revelou que a fuga de informação por *Attribute Inference* é fortemente concentrada na Profissao e no TotOutflowU3, que apresentam ganhos de 0.207 e 0.157 respetivamente, enquanto variáveis como RacSaldoMedioMinU6 e FlagVencU12 têm fuga de informação negligenciável. Esta assimetria sugere uma abordagem diferente: calibrar a perturbação pela fuga de informação efetiva de cada variável, concentrando o ruído nas variáveis problemáticas e mantendo intactas as restantes.

Esta é a hipótese central da perturbação seletiva, avaliada nesta secção através de dois métodos desenvolvidos como extensão natural dos métodos uniformes testados anteriormente.

4.6.1 RCS Seletivo

O *Random Category Substitution* Seletivo (RCS Seletivo) é uma extensão do RCS que calibra a intensidade da perturbação em função da fuga de informação efetiva de cada variável. As 14 variáveis são agrupadas em três níveis: maior fuga de informação (Profissao, TotOutflowU3, Prazo, Idade), fuga de informação intermédia (TipoCC, Finalidade, AntBanco, DifNumCredSFDtDt3, RacMontanteExpTot) e menor fuga de informação (ValorVencMedioSFU12, ExpSFDt6, NumCredSFDt6, RacSaldoMedioMinU6, FlagVencU12). A cada grupo é atribuída uma probabilidade de perturbação distinta, p_{high}, p_{mid} e p_{low}, permitindo concentrar o ruído nas variáveis problemáticas e preservar a utilidade nas restantes.

O agrupamento foi definido com base nos valores de *Attribute Inference* do modelo de referência, descritos na Secção 4.1. As variáveis de maior fuga de informação apresentam ganhos superiores a 0.10, as de fuga intermédia entre 0.05 e 0.10, e as de menor fuga abaixo de 0.05.

Para identificar a melhor combinação de parâmetros, foi realizada uma *grid search* sobre o espaço de configurações possíveis. O parâmetro p_{high} foi variado entre 0.05 e 0.70, cobrindo o espectro desde uma perturbação ligeira até uma perturbação intensa nas variáveis mais problemáticas. O parâmetro p_{mid} foi variado entre 0.00 e 0.30, permitindo testar desde a ausência total de perturbação nas variáveis de fuga intermédia até uma perturbação moderada. O parâmetro p_{low} foi testado com três valores fixos (0.00, 0.05 e 0.10), dado que as variáveis de menor fuga de

informação têm contributo negligenciável e não justificam perturbação elevada. Os resultados foram avaliados para os dois modelos, filtrando as configurações com AUC-ROC superior a 0.75 e ordenando pela soma do *Attribute Inference* nas variáveis de maior fuga de informação.

A Figura 4.2 ilustra o espaço de pesquisa para os dois modelos. Cada ponto representa uma configuração testada, com o eixo horizontal a mostrar o AUC-ROC obtido e o eixo vertical a mostrar a soma do *Attribute Inference* nas variáveis de maior fuga de informação. A cor de cada ponto indica o valor de p_{high} utilizado. O modelo de referência está assinalado com uma estrela e a melhor configuração encontrada com um losango vermelho.

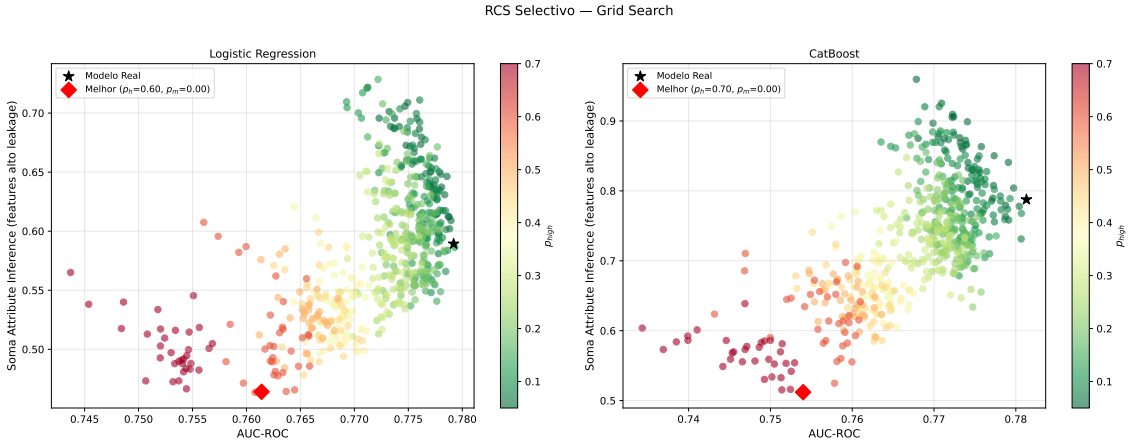

Figura 4.2: *Grid search* RCS Seletivo — espaço de pesquisa e melhor configuração

A figura evidencia o compromisso entre utilidade e privacidade: configurações com p_{high} elevado (tons verdes) tendem a reduzir a soma do *Attribute Inference* mas também o AUC-ROC, enquanto configurações com p_{high} baixo (tons cor-de-rosa) preservam melhor a utilidade mas com menor redução da fuga de informação. A melhor configuração encontrada situa-se no canto inferior esquerdo do espaço de soluções aceitáveis, a que melhor equilibra os dois objetivos dentro do limiar de AUC-ROC de 0.75.

A Tabela 4.24 apresenta as métricas de utilidade e privacidade global para as melhores configurações encontradas.

Tabela 4.24: Métricas de utilidade e privacidade global — RCS Seletivo

Config.	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
LR p_h=0.60 p_m=0.00	0.761	0.407	0.696	0.625	10.43
LR p_h=0.60 p_m=0.05	0.764	0.415	0.706	0.632	9.02
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
CAT p_h=0.70 p_m=0.00	0.754	0.382	0.685	0.657	1.38
CAT p_h=0.60 p_m=0.03	0.758	0.388	0.683	0.639	1.12

Os resultados confirmam a hipótese da perturbação seletiva. O RCS Seletivo LR com $p_{high} = 0.60$ reduz a soma do *Attribute Inference* nas variáveis de maior fuga de informação de 0.589

para 0.464, uma redução de 21.2%, com AUC-ROC de 0.761. A Profissao reduz de 0.207 para 0.140 (-32%) e o Prazo de 0.108 para 0.049 (-55%). O TotOutflowU3 mantém-se praticamente inalterado, pois a sua correlação com a variável alvo é mais resistente à perturbação.

O *CatBoost* com $p_{high} = 0.70$ tem a maior redução absoluta na soma (de 0.788 para 0.512, -35%), com Prazo a descer para 0.011, praticamente eliminado. No entanto, o MIA sobe para 0.657 e o Canary Gap mantém-se baixo, confirmando a tendência de memorização do *CatBoost*.

A Tabela 4.25 apresenta o *Attribute Inference* nas variáveis de maior fuga de informação para as melhores configurações.

Tabela 4.25: *Attribute Inference* — RCS Seletivo

Config.	**Profissao**	**TotOutflow**	**Prazo**	**Idade**	**Soma**
LR (Ref.)	0.207	0.157	0.108	0.119	0.589
LR p_h=0.60 p_m=0.00	0.140	0.157	0.049	0.119	0.464
LR p_h=0.60 p_m=0.05	0.143	0.155	0.049	0.117	0.464
CAT (Ref.)	0.251	0.231	0.148	0.158	0.788
CAT p_h=0.70 p_m=0.00	0.182	0.159	0.011	0.159	0.512
CAT p_h=0.60 p_m=0.03	0.212	0.166	0.001	0.146	0.525

O MIA elevado em todas as configurações seletivas reflete o desfasamento de distribuição entre dados perturbados de treino e dados reais de teste, o mesmo fenómeno observado em todos os métodos de perturbação anteriores. O RCS Seletivo demonstra que a perturbação calibrada pela fuga de informação efetiva de cada variável consegue resultados mais equilibrados do que a perturbação uniforme. No entanto, a ausência de garantias formais de privacidade levantou a questão de se seria possível obter resultados semelhantes com garantias teóricas, motivando a avaliação do GRR Seletivo na secção seguinte.

4.6.2 GRR Seletivo

O *Generalized Randomized Response* Seletivo (GRR Seletivo) combina a abordagem seletiva desenvolvida no RCS Seletivo com as garantias formais de privacidade do GRR. Em vez de um único ε aplicado a todas as variáveis, são definidos valores distintos por grupo de fuga de informação: ε baixo (mais privacidade) para as variáveis de maior fuga de informação, ε médio para as de fuga de informação intermédia, e sem perturbação para as de menor fuga de informação. Esta abordagem mantém a garantia formal ε-LDP por variável, com um orçamento de privacidade diferenciado em função da sensibilidade de cada grupo.

O agrupamento de variáveis é o mesmo utilizado no RCS Seletivo. A perturbação é aplicada diretamente sobre as variáveis categóricas antes da codificação binária multivariada, tal como no GRR uniforme.

Para identificar a melhor combinação de parâmetros, foi realizada uma *grid search* sobre o espaço de configurações possíveis. O parâmetro ε_{high} foi variado entre 0.5 e 7.0, cobrindo o espectro desde uma privacidade muito forte até uma privacidade moderada nas variáveis mais

problemáticas. O parâmetro ε_{mid} foi variado entre 1.0 e 7.0, testando diferentes níveis de proteção para as variáveis de fuga de informação intermédia. As variáveis de menor fuga de informação não foram perturbadas em nenhuma configuração. Os resultados foram filtrados por AUC-ROC superior a 0.74 e ordenados pela soma do *Attribute Inference* nas variáveis de maior fuga de informação.

A Figura 4.3 ilustra o espaço de pesquisa para os dois modelos. Cada ponto representa uma configuração testada, com o eixo horizontal a mostrar o AUC-ROC obtido e o eixo vertical a mostrar a soma do *Attribute Inference* nas variáveis de maior fuga de informação. A cor de cada ponto indica o valor de ε_{high} utilizado. O modelo de referência está assinalado com uma estrela e a melhor configuração encontrada com um losango vermelho.

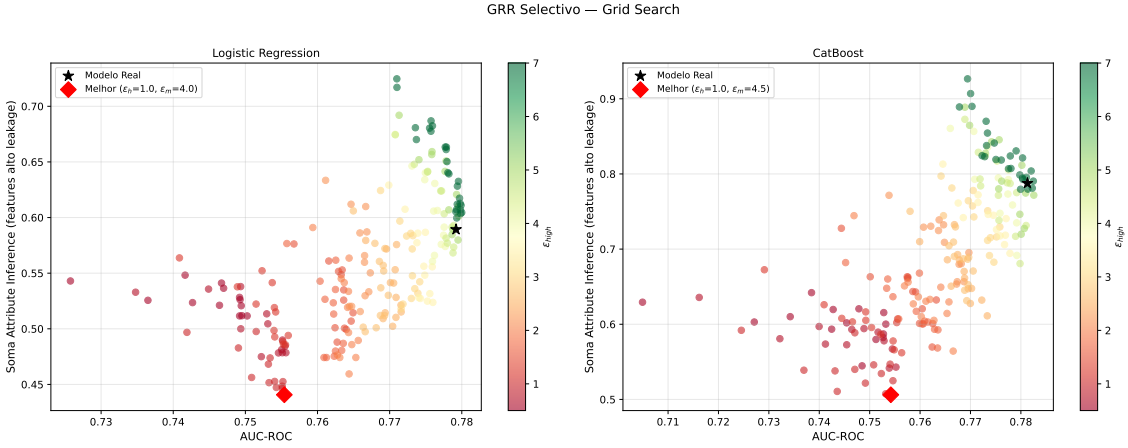

Figura 4.3: *Grid search* GRR Seletivo — espaço de pesquisa e melhor configuração

A figura evidencia o mesmo padrão observado no RCS Seletivo: valores de ε_{high} mais baixos (mais privacidade, tons cor-de-rosa) tendem a reduzir mais a fuga de informação mas também o AUC-ROC, enquanto valores mais altos (menos privacidade, tons verdes) preservam melhor a utilidade com menor redução da fuga de informação.

A Tabela 4.26 apresenta as métricas de utilidade e privacidade global para as melhores configurações encontradas.

Tabela 4.26: Métricas de utilidade e privacidade global — GRR Seletivo

Config.	**AUC-ROC**	**KS**	**G-Mean**	**MIA**	**Canary Gap**
LR (Ref.)	0.7792	0.4358	0.7138	0.4986	19.49
LR ε_h=1.0 ε_m=4.0	0.755	0.399	0.696	0.634	7.51
LR ε_h=1.0 ε_m=5.0	0.755	0.400	0.696	0.632	8.51
LR ε_h=2.0 ε_m=5.0	0.764	0.411	0.697	0.580	8.98
CAT (Ref.)	0.7813	0.4361	0.7123	0.5032	1.90
CAT ε_h=1.0 ε_m=4.5	0.754	0.383	0.680	0.633	1.59
CAT ε_h=1.0 ε_m=5.0	0.753	0.379	0.681	0.633	3.11

A Tabela 4.27 apresenta o *Attribute Inference* nas variáveis de maior fuga de informação para as melhores configurações.

Tabela 4.27: *Attribute Inference* — GRR Seletivo

Config.	**Profissao**	**TotOutflow**	**Prazo**	**Idade**	**Soma**
LR (Ref.)	0.207	0.157	0.108	0.119	0.589
LR ε_h=1.0 ε_m=4.0	0.118	0.138	0.070	0.114	0.441
LR ε_h=1.0 ε_m=5.0	0.118	0.146	0.066	0.117	0.446
LR ε_h=2.0 ε_m=5.0	0.143	0.137	0.065	0.113	0.459
CAT (Ref.)	0.251	0.231	0.148	0.158	0.788
CAT ε_h=1.0 ε_m=4.5	0.137	0.183	0.066	0.120	0.506
CAT ε_h=1.0 ε_m=5.0	0.118	0.175	0.083	0.132	0.507

O GRR Seletivo é o método com melhor redução simultânea da fuga de informação com garantia formal de privacidade de todos os testados. A configuração LR com $\varepsilon_{high} = 1.0$ e $\varepsilon_{mid} = 4.0$ reduz a soma do *Attribute Inference* de 0.589 para 0.441, uma redução de 25.1%, com AUC-ROC de 0.755. A Profissao reduz de 0.207 para 0.118 (-43%) e o TotOutflowU3 de 0.157 para 0.138 (-12%), sendo a única configuração em toda a experimentação que reduz simultaneamente estas duas variáveis com garantia formal de privacidade. No *CatBoost*, a configuração com $\varepsilon_{high} = 1.0$ e $\varepsilon_{mid} = 4.5$ alcança uma redução de 35.8% na soma do *Attribute Inference* (de 0.788 para 0.506), com AUC-ROC de 0.754, a maior redução absoluta de todos os métodos testados.

Comparando com o RCS Seletivo, o GRR Seletivo apresenta melhor redução da Profissao e do TotOutflowU3 mas AUC-ROC ligeiramente inferior (0.755 vs 0.761 no LR e 0.754 vs 0.754 no *CatBoost*). A vantagem determinante do GRR Seletivo é precisamente a garantia formal ε-LDP com $\varepsilon = 1.0$ para as variáveis de maior fuga de informação, garantia que o RCS Seletivo não oferece.

A configuração LR com $\varepsilon_{high} = 2.0$ e $\varepsilon_{mid} = 5.0$ apresenta um compromisso diferente (AUC-ROC de 0.764, MIA de 0.580 e soma de 0.459), com menor degradação de utilidade para quem priorize o desempenho preditivo em detrimento de garantias de privacidade mais fortes.

Os resultados do GRR Seletivo encerram a análise experimental. O capítulo seguinte apresenta uma discussão comparativa de todos os métodos avaliados, procurando identificar os padrões transversais e as conclusões centrais desta investigação.

Capítulo 5

Discussão e Conclusões

A experimentação desenvolvida ao longo do capítulo anterior cobriu um espectro alargado de técnicas de PPML, organizadas em cinco famílias distintas: mecanismos de Privacidade Diferencial Local, geração de dados sintéticos com Privacidade Diferencial, proteção do modelo durante o treino, perturbação por ruído simples e perturbação seletiva. Este capítulo sintetiza os resultados obtidos, discutindo os padrões transversais entre famílias de métodos, o compromisso entre privacidade e utilidade, a arquitetura integrada proposta para adoção pela ITSCredit, as limitações do trabalho e as direções de investigação futura.

5.1 Análise Comparativa dos Métodos

Os resultados obtidos ao longo da experimentação evidenciam padrões consistentes entre famílias de métodos. A fuga de informação por inferência de atributos revelou-se estrutural e proporcional ao poder preditivo de cada variável: qualquer método que preserve a utilidade do modelo preserva inevitavelmente as correlações que estão na origem da fuga de informação. Os padrões observados em cada família de métodos são resumidos de seguida:

- **Mecanismos de Privacidade Diferencial Local** — O GRR, OUE e RAPPOR degradam a utilidade de forma acentuada para níveis de privacidade forte, e o MIA sobe em todos os casos por desfasamento de distribuição. O GRR Médio é o melhor desta família, com redução da Profissao para 0.177, mas sem resolver o TotOutflowU3 e com AUC-ROC de 0.744.

- **Dados Sintéticos com Privacidade Diferencial** — O DP-CTGAN e a DPGAN falharam por incompatibilidade estrutural com as características do conjunto de dados, nomeadamente desbalanceamento extremo e ausência de geração condicional. O DP FT SD mantém utilidade razoável mas não melhora o *Attribute Inference*, preservando as correlações estruturais que causam a fuga de informação.

- **Proteção do Modelo** — A Privacidade Diferencial no treino do classificador é incompatível com dados de codificação binária multivariada. O PATE protege eficazmente contra

memorização mas não resolve correlações populacionais. A aprendizagem com restrições de equidade amplifica paradoxalmente a fuga de informação das variáveis que protege, confirmando que equidade e privacidade são objetivos distintos e potencialmente contraditórios [62].

- **Perturbação por Ruído Simples** — Os métodos de perturbação simples mantêm boa utilidade para valores baixos de perturbação mas não reduzem a fuga de informação de forma consistente — a perturbação uniforme não discrimina entre variáveis problemáticas e não problemáticas.

- **Perturbação Seletiva** — A calibração da perturbação pela fuga de informação efetiva de cada variável revelou-se a abordagem mais eficaz. O RCS Seletivo reduz a soma do *Attribute Inference* nas variáveis de maior fuga de informação em 21.2% no LR e 35.0% no CatBoost, mas sem garantias formais de privacidade. O GRR Seletivo é o único método que reduz simultaneamente a Profissao e o TotOutflowU3 com garantia formal ε-LDP, alcançando uma redução de 25.1% no LR e 35.8% no CatBoost, com $\varepsilon = 1.0$ para as variáveis de maior fuga de informação.

A Figura 5.1 apresenta o *Attribute Inference* nas quatro variáveis de maior fuga de informação para os métodos mais representativos de cada família, permitindo uma comparação direta do impacto de cada abordagem. O heatmap principal mostra o valor absoluto de fuga de informação por variável e por método, com cores que variam do vermelho escuro (fuga de informação elevada) ao verde escuro (fuga de informação baixa). A coluna $\Delta(\%)$ indica a variação percentual na soma das quatro variáveis face ao modelo de referência, com verde a indicar melhoria e vermelho a indicar agravamento.

O heatmap evidencia com clareza os padrões identificados ao longo da experimentação. O RCS p=0.05, representativo da perturbação simples uniforme, praticamente não altera a fuga de informação face ao modelo de referência (Δ=+1.6% no LR e +1.3% no CatBoost), e a degradação de utilidade é negligenciável (AUC-ROC de 0.7796 no LR e 0.7787 no CatBoost, uma perda de apenas 0.000 e 0.003 respetivamente). O GRR Médio consegue alguma redução (Δ=$-$2.8% no LR e $-$23.0% no CatBoost), mas à custa de uma degradação de utilidade significativa: AUC-ROC de 0.7437 no LR e 0.7322 no CatBoost, uma perda de 0.036 e 0.049 respetivamente, e sem resolver a Profissao e o TotOutflowU3 de forma consistente em ambos os modelos.

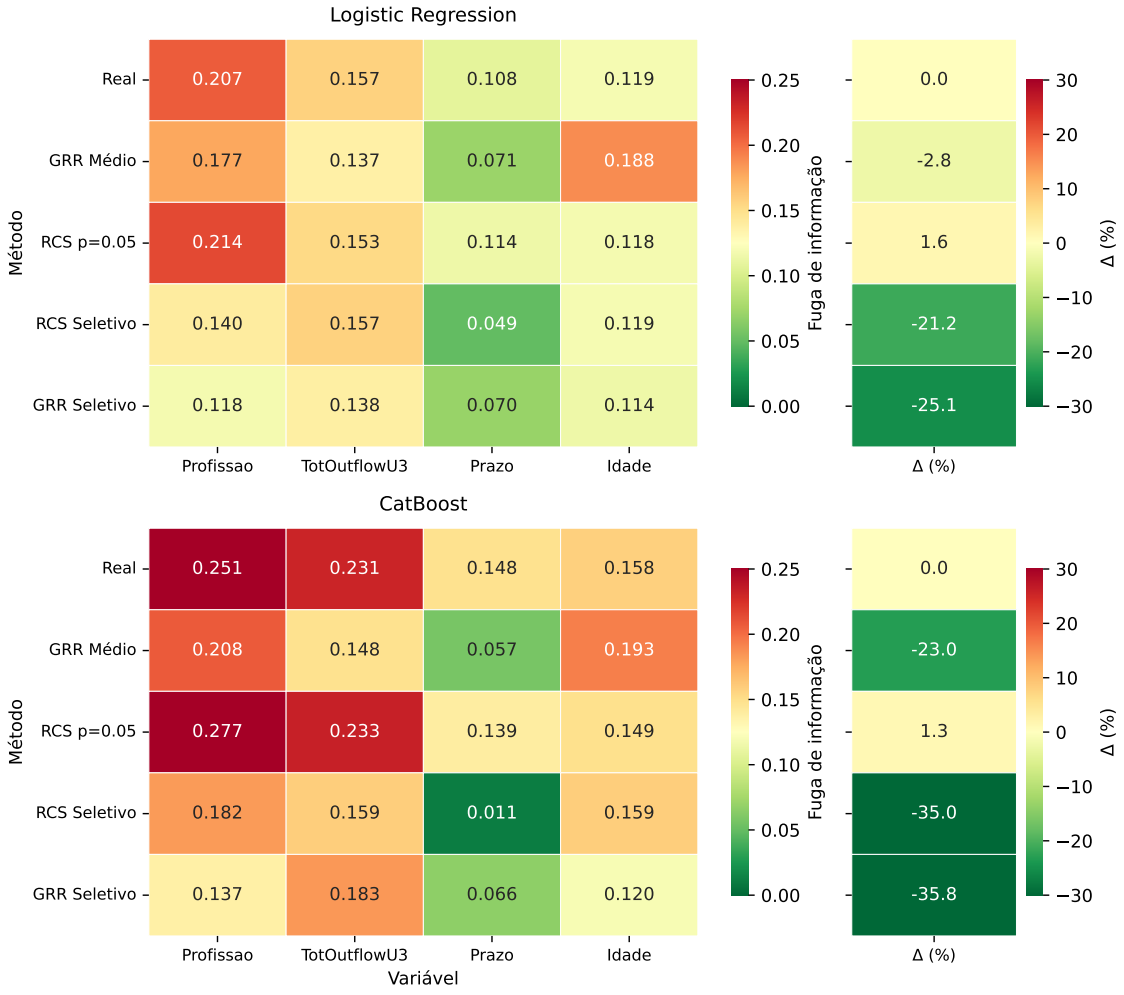

Figura 5.1: *Attribute Inference* por método e variável — Logistic Regression e CatBoost

A perturbação seletiva destaca-se visivelmente no heatmap. O RCS Seletivo reduz a fuga de informação de forma mais equilibrada (Δ=−21.2% no LR e −35.0% no CatBoost), com a Profissao a descer para 0.140 no LR e o Prazo praticamente eliminado no CatBoost (0.011), com uma degradação de utilidade moderada: AUC-ROC de 0.761 no LR e 0.754 no CatBoost, uma perda de 0.018 e 0.027 respetivamente. O GRR Seletivo alcança a maior redução global em ambos os modelos (Δ=−25.1% no LR e −35.8% no CatBoost), sendo visualmente o método com células mais verdes no heatmap, e é o único que reduz simultaneamente a Profissao e o TotOutflowU3 com garantia formal ε-LDP, com uma degradação de utilidade semelhante ao RCS Seletivo: AUC-ROC de 0.755 no LR e 0.754 no CatBoost, uma perda de 0.024 e 0.027 respetivamente.

5.2 Compromisso Privacidade-Utilidade

O compromisso entre privacidade e utilidade é um tema central na literatura de PPML [56], mas os resultados obtidos neste trabalho revelam uma dimensão adicional que vai além do compromisso clássico, nomeadamente a fuga de informação estrutural.

Em todos os métodos testados, observa-se o mesmo padrão: a redução do *Attribute Inference* nas variáveis mais problemáticas (Profissao e TotOutflowU3) implica necessariamente uma degradação de utilidade. Esta é uma consequência direta da natureza dos dados: Profissao e TotOutflowU3 têm fuga de informação elevada precisamente porque são variáveis preditivas da variável alvo. Um modelo com boa capacidade discriminativa aprende as correlações entre estas variáveis e o risco de incumprimento, e é exatamente essa informação que o *Attribute Inference* explora.

Esta relação entre fuga de informação e poder preditivo define um limite para a redução de privacidade que qualquer método pode alcançar sem sacrificar utilidade. Não existe método que elimine completamente o *Attribute Inference* nas variáveis mais preditivas sem tornar o modelo inutilizável, porque eliminar a fuga de informação implica eliminar a informação que o modelo usa para discriminar.

O MIA elevado observado em todos os métodos de perturbação é um problema distinto: não reflete memorização de exemplos individuais, mas sim o desfasamento entre a distribuição dos dados de treino perturbados e os dados de teste reais. Este fenómeno foi confirmado empiricamente: o MIA calculado sobre dados reais mantém-se próximo de 0.50 em todas as configurações, ou seja, o modelo perturbado não memoriza, mas os dados perturbados são estatisticamente distinguíveis dos dados reais.

O GRR Seletivo com $\varepsilon_{high} = 1.0$ representa o melhor compromisso encontrado. No LR, a configuração com $\varepsilon_{mid} = 4.0$ alcança uma redução de 25.1% na soma do *Attribute Inference* nas variáveis de maior fuga de informação, com AUC-ROC de 0.755 e garantia formal ε-LDP. No *CatBoost*, a configuração com $\varepsilon_{mid} = 4.5$ alcança uma redução de 35.8%, com AUC-ROC de 0.754, a maior redução absoluta de todos os métodos testados em ambos os modelos. A configuração LR com $\varepsilon_{high} = 2.0$ e $\varepsilon_{mid} = 5.0$ oferece um compromisso alternativo (AUC-ROC de 0.764 com menor redução de fuga de informação), adequado para contextos onde a utilidade é prioritária.

Em termos práticos, para um sistema de avaliação de risco de crédito em produção, a escolha entre estas configurações depende dos requisitos regulatórios e de negócio. Se a garantia formal de privacidade é exigida, o GRR Seletivo com $\varepsilon_{high} = 1.0$ é a escolha mais defensável em ambos os modelos. Se a utilidade é prioritária e a garantia formal não é obrigatória, o RCS Seletivo com $p_{high} = 0.60$ oferece AUC-ROC mais alto com redução comparável do *Attribute Inference*: 21.2% no LR e 35.0% no *CatBoost*.

5.3 Arquitetura Integrada e Recomendação

Os resultados experimentais obtidos permitem responder de forma concreta à questão central desta dissertação: qual a abordagem mais adequada para proteger a privacidade num sistema real de avaliação de risco de crédito, mantendo a utilidade preditiva necessária para a tomada de decisão? A resposta emerge com clareza da análise comparativa: o *GRR Seletivo* com $\varepsilon_{high} = 1.0$ e $\varepsilon_{mid} = 4.5$ é a abordagem mais adequada para adoção pela ITSCredit, sendo o único método que combina, em simultâneo, redução efetiva da fuga de informação nas variáveis mais sensíveis, garantia formal de privacidade diferencial local e preservação de utilidade preditiva a um nível

operacionalmente aceitável. Recomenda-se a configuração CatBoost com $\varepsilon_{high} = 1.0$ e $\varepsilon_{mid} = 4.5$, que alcança uma redução de 35.8% na soma da fuga de informação nas quatro variáveis de maior fuga de informação com AUC-ROC de 0.754. O CatBoost é preferível à Regressão Logística neste contexto por apresentar maior redução absoluta da fuga de informação com utilidade equivalente.

A Figura 5.2 ilustra a arquitetura integrada proposta, que incorpora o GRR Seletivo no fluxo de processamento existente da ITSCredit.

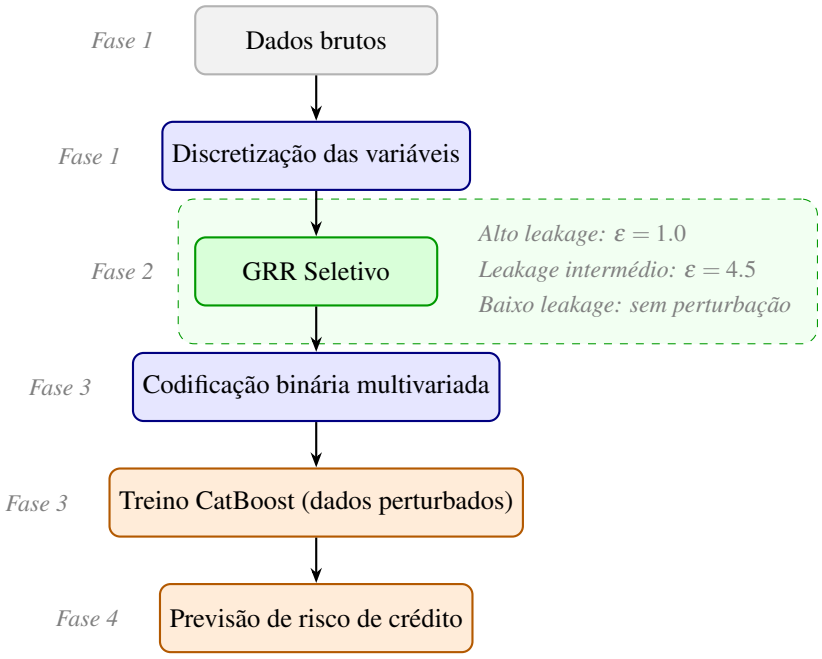

Figura 5.2: Arquitetura integrada proposta — fluxo de processamento com GRR Seletivo

O fluxo funciona em quatro fases. Na primeira fase, os dados brutos do cliente são recolhidos e as variáveis são discretizadas segundo os intervalos definidos na metodologia. Esta fase é idêntica ao fluxo atual e não é alterada pela introdução do mecanismo de privacidade. Na segunda fase, o GRR Seletivo é aplicado sobre as variáveis discretizadas antes da codificação: as variáveis de maior fuga de informação (Profissao e TotOutflowU3) são perturbadas com $\varepsilon = 1.0$, as de fuga intermédia com $\varepsilon = 4.5$, e as de menor fuga não são perturbadas. Na terceira fase, as variáveis perturbadas são codificadas em formato binário multivariado e o classificador CatBoost é treinado sobre estes dados, aprendendo diretamente sobre os dados perturbados, sem nunca ter acesso aos valores originais das variáveis sensíveis. Na quarta fase, em produção, os dados de novos clientes são perturbados com o mesmo mecanismo antes de serem passados ao modelo, garantindo que as previsões são sempre geradas a partir de dados com garantia formal ε-LDP.

Esta arquitetura tem três propriedades relevantes para o contexto da ITSCredit. Primeiro, a perturbação é aplicada antes do treino do modelo, o que significa que o modelo nunca tem acesso aos valores originais das variáveis mais sensíveis, pois a privacidade resulta da própria estrutura do fluxo de processamento e não depende de restrições de acesso impostas posteriormente. Segundo, a garantia ε-LDP é formal e verificável, o que facilita a justificação perante requisitos regulatórios

como o RGPD. Terceiro, a degradação de utilidade é controlada e aceitável: a AUC-ROC de 0.755 no LR e 0.754 no CatBoost representa uma perda de apenas 0.024 e 0.027 face ao modelo de referência, mantendo o poder discriminativo necessário para a avaliação de risco.

A adoção desta arquitetura representa um passo concreto e tecnicamente fundamentado para a proteção da privacidade nos sistemas de decisão de crédito da ITSCredit, conciliando os requisitos operacionais de desempenho preditivo com as obrigações regulatórias e éticas de proteção de dados dos clientes.

5.4 Limitações do Trabalho

Os resultados obtidos neste trabalho devem ser interpretados à luz de um conjunto de limitações inerentes ao contexto em que a investigação foi conduzida.

A primeira limitação diz respeito à especificidade do conjunto de dados utilizado. A experimentação foi conduzida sobre um único conjunto de dados de crédito, com características muito específicas: estrutura tabular com discretização, seleção reduzida de variáveis por requisitos do domínio e codificação por pesos de evidência. Os resultados obtidos refletem estas especificidades e a sua generalização para outros contextos bancários ou outros tipos de produto de crédito não está garantida. Em particular, fluxos de processamento sem discretização ou com variáveis contínuas podem apresentar comportamentos distintos face aos métodos testados.

A segunda limitação é o desbalanceamento extremo da variável alvo. A taxa de incumprimento de 4.4% condiciona de forma determinante a aplicabilidade de métodos de síntese de dados e de treino com Privacidade Diferencial, como ficou demonstrado com o DP-CTGAN e a DPGAN. Este desbalanceamento é característico de carteiras de crédito reais e representa uma restrição estrutural do problema, não uma limitação metodológica do trabalho.

A terceira limitação decorre da obrigatoriedade de discretização de todas as variáveis antes do treino do modelo, por requisito do domínio. Esta etapa condiciona de duas formas o espaço de métodos de PPML aplicáveis: por um lado, a perturbação aplicada antes da discretização sobre variáveis contínuas pode ser parcialmente absorvida ou distorcida pelo processo de discretização subsequente, tornando difícil controlar o nível de privacidade efetivamente alcançado; por outro lado, a perturbação aplicada após a discretização fica restrita a mecanismos desenhados para dados categóricos, excluindo mecanismos formulados para variáveis contínuas. Esta tensão, inerente ao fluxo de processamento utilizado, limita a comparabilidade dos métodos testados com abordagens avaliadas na literatura em contextos sem discretização.

A quarta limitação decorre da utilização de dados reais. A complexidade e sensibilidade dos dados de crédito amplificam os problemas de todos os métodos testados face ao que seria observado em conjuntos de dados públicos de referência. As correlações fortes entre variáveis e variável alvo, a presença de variáveis demograficamente sensíveis e as restrições regulatórias do sector tornam este problema estruturalmente mais difícil do que os cenários experimentais tipicamente avaliados na literatura de PPML.

Por fim, a fuga de informação estrutural identificada neste trabalho impõe um limite teórico à redução de privacidade alcançável sem sacrificar utilidade. Enquanto as variáveis mais problemáticas forem simultaneamente as mais preditivas, qualquer método de PPML enfrenta este compromisso fundamental: reduzir a fuga de informação implica reduzir o sinal preditivo, e vice-versa.

5.5 Trabalho Futuro

Os resultados obtidos neste trabalho definem várias direções de investigação futura:

- A experimentação foi conduzida sobre um único conjunto de dados de crédito, com características muito específicas em termos de estrutura, distribuição e domínio. A validação dos resultados noutros contextos (diferentes produtos financeiros, diferentes mercados ou diferentes estruturas de dados) é uma direção natural para avaliar em que medida os padrões identificados se generalizam, nomeadamente a relação entre fuga de informação estrutural e poder preditivo das variáveis.

- A avaliação foi limitada a dois modelos (Regressão Logística e CatBoost). A natureza discreta dos dados, resultante da obrigatoriedade de discretização, condiciona também a família de modelos aplicáveis: modelos como redes neuronais são tipicamente concebidos para dados contínuos e podem não ser adequados neste contexto sem transformações adicionais. Ainda assim, outros modelos compatíveis com dados tabulares categóricos, como árvores de decisão ou modelos de gradiente extremo com diferentes configurações, podem apresentar comportamentos distintos face aos métodos de perturbação testados, nomeadamente no que diz respeito à fuga de informação estrutural e à resistência ao desfasamento de distribuição introduzido pela perturbação.

- O agrupamento de variáveis utilizado no GRR Seletivo e no RCS Seletivo foi definido manualmente com base nos valores de *Attribute Inference* do modelo de referência. Uma direção promissora seria desenvolver um método de agrupamento automático que determine os grupos e os parâmetros de perturbação ótimos para um dado limiar de AUC-ROC, eliminando a necessidade de intervenção manual e tornando a abordagem mais facilmente aplicável a novos conjuntos de dados.

- Os conjuntos de dados de crédito têm uma dimensão temporal relevante: as distribuições de risco evoluem ao longo do tempo em resposta a ciclos económicos, alterações regulatórias e mudanças no comportamento dos clientes. Avaliar o comportamento dos métodos de perturbação seletiva em cenários de re-treino periódico do modelo é uma direção importante para a aplicação prática em produção.

- A perturbação seletiva calibrada pela fuga de informação pode introduzir desequilíbrios nas previsões para grupos demográficos específicos, pois ao perturbar mais intensamente

variáveis como Profissao e Idade, que têm correlação com características demográficas protegidas, o método pode afetar de forma diferenciada grupos distintos. Avaliar o impacto da perturbação seletiva em métricas de equidade algorítmica é uma direção relevante, dado o crescente enquadramento regulatório nesta matéria no sector financeiro.

- Por fim, as métricas utilizadas neste trabalho (AUC-ROC, KS, MIA e *Attribute Inference*) são métricas técnicas que permitem comparar métodos de forma rigorosa mas abstrata. Uma direção de trabalho futuro com elevado valor prático é traduzir o impacto dos métodos de perturbação seletiva em termos concretos para as tabelas de pontuação de crédito utilizadas em produção, avaliando como a perturbação afeta os intervalos de pontuação, as taxas de aprovação e a discriminação entre segmentos de risco.

Bibliografia

[1] Lyn C. Thomas, David B. Edelman e Jonathan N. Crook. *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford University Press, 2009.

[2] Stefan Lessmann et al. «Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research». Em: *European Journal of Operational Research* 247.1 (2015), pp. 124–136. DOI: 10.1016/j.ejor.2015.05.030. URL: https://doi.org/10.1016/j.ejor.2015.05.030.

[3] Huyen Giang Thi Thu, Thang Viet Doan e Tai Le Quy. «An experimental study on fairness-aware machine learning for credit scoring problem». Em: *arXiv preprint arXiv:2412.20298* (2024). URL: https://arxiv.org/abs/2412.20298.

[4] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)*. Official Journal of the European Union, L 119, 4 May 2016. CELEX: 32016R0679. ELI: http://data.europa.eu/eli/reg/2016/679/oj. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

[5] European Banking Authority. *Guidelines on loan origination and monitoring*. 2020. URL: https://www.eba.europa.eu/regulation-and-policy/credit-risk/guidelines-on-loan-origination-and-monitoring.

[6] Runhua Xu, Nathalie Baracaldo e James Joshi. «Privacy-Preserving Machine Learning: Methods, Challenges and Directions». Em: *arXiv preprint arXiv:2108.04417* (2021). URL: https://arxiv.org/abs/2108.04417.

[7] Peter Kairouz et al. «Advances and Open Problems in Federated Learning». Em: *Foundations and Trends in Machine Learning* 14.1–2 (2021), pp. 1–210. DOI: 10.1561/2200000083. URL: https://doi.org/10.1561/2200000083.

[8] Lyn C. Thomas, David B. Edelman e Jonathan N. Crook. *Credit Scoring and Its Applications*. SIAM, 2002. ISBN: 9780898718317.

[9] Francisco Louzada, Anderson Ara e Guilherme B. Fernandes. «Classification methods applied to credit scoring: Systematic review and overall comparison». Em: *Surveys in Operations Research and Management Science* 21.2 (2016), pp. 117–134. DOI: 10.1016/j.sorms.2016.10.001. URL: https://arxiv.org/abs/1602.02137.

[10] European Central Bank. *ECB Guide to Internal Models*. ECB Banking Supervision. Revised version published 19 February 2024. 2024. URL: https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guides202402_internalmodels.en.pdf.

[11] Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, 2012. ISBN: 9781119279150.

[12] Xolani P. Dastile, Turgay Celik e Mmutlanyane Potsane. «Statistical and Machine Learning models in credit scoring: A systematic literature survey». Em: *Applied Soft Computing* 91 (2020), p. 106263. DOI: 10.1016/j.asoc.2020.106263. URL: https://doi.org/10.1016/j.asoc.2020.106263.

[13] Helmi Ayari, Ramzi Guetari e Naoufel Kraïem. «Machine Learning powered financial credit scoring: A systematic literature review of methods and trends (2018–2024)». Em: *Artificial Intelligence Review* 59.1 (2025), p. 13. DOI: 10.1007/s10462-025-11416-2. URL: https://link.springer.com/article/10.1007/s10462-025-11416-2.

[14] Basel Committee on Banking Supervision. *Principles for the Management of Credit Risk*. Bank for International Settlements. BCBS publication d595, published 30 April 2025. 2025. URL: https://www.bis.org/bcbs/publ/d595.pdf.

[15] Information Commissioner's Office. *A Guide to the Data Protection Principles*. 2023. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/.

[16] European Banking Authority. *Final Report – Guidelines on loan origination and monitoring*. EBA/GL/2020/06. 2020. URL: https://www.eba.europa.eu/sites/default/files/document_library/Publications/Guidelines/2020/Guidelines%20on%20loan%20origination%20and%20monitoring/884283/EBA%20GL%202020%2006%20Final%20Report%20on%20GL%20on%20loan%20origination%20and%20monitoring.pdf.

[17] European Data Protection Supervisor. *Opinion 11/2021 on the Proposal for a Directive on Consumer Credits*. Adopted 26 August 2021. 2021. URL: https://www.edps.europa.eu/system/files/2021-08/opinion_consumercredit-final_en.pdf.

[18] Comissão Nacional de Proteção de Dados. *Guia do Comité Europeu apresenta metodologia para gerir riscos de privacidade em sistemas de IA*. Notícia CNPD (referencia documento do EDPB). 2025. URL: https://www.cnpd.pt/comunicacao-publica/noticias/guia-do-comite-europeu-apresenta-metodologia-para-gerir-riscos-de-privacidade-em-sistemas-de-ia/.

[19] European Data Protection Board. *AI Privacy Risks & Mitigations – Large Language Models (LLMs)*. EDPB Support Pool of Experts. 2025. URL: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf.

[20] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, L, 2024/1689, 12 July 2024. CELEX: 32024R1689. ELI: http://data.europa.eu/eli/reg/2024/1689/oj. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[21] Simon Henseler. *Data Protection in Credit Scoring by Credit Information Agencies*. SwissPrivacy.law (resumo de tese de doutoramento). 2025. URL: https://swissprivacy.law/376/ (acedido em 04/02/2026).

[22] Hongsheng Hu et al. «Membership Inference Attacks on Machine Learning: A Survey». Em: *ACM Computing Surveys* 54.11s (2022), pp. 1–37. DOI: 10.1145/3523273. URL: https://doi.org/10.1145/3523273.

[23] Nicholas Carlini et al. «Extracting Training Data from Large Language Models». Em: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2633–2650. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[24] Reza Shokri et al. «Membership Inference Attacks Against Machine Learning Models». Em: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18. DOI: 10.1109/SP.2017.41. URL: https://doi.org/10.1109/SP.2017.41.

[25] Matt Fredrikson, Somesh Jha e Thomas Ristenpart. «Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures». Em: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677. URL: https://doi.org/10.1145/2810103.2813677.

[26] Samuel Yeom et al. «Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting». Em: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. 2018, pp. 268–282. DOI: 10.1109/CSF.2018.00027. URL: https://arxiv.org/abs/1709.01604.

[27] Nicholas Carlini et al. «The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks». Em: *28th USENIX Security Symposium (USENIX Security 19)*. 2019. URL: https://www.usenix.org/system/files/sec19-carlini.pdf.

[28] Latanya Sweeney. «k-Anonymity: A Model for Protecting Privacy». Em: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (2002), pp. 557–570. DOI: 10.1142/S0218488502001648. URL: https://doi.org/10.1142/S0218488502001648.

[29] Ashwin Machanavajjhala et al. «l-Diversity: Privacy Beyond k-Anonymity». Em: *ACM Transactions on Knowledge Discovery from Data* 1.1 (2007), p. 3. DOI: 10.1145/1217299.1217302. URL: https://doi.org/10.1145/1217299.1217302.

[30] Bo Liu et al. «When Machine Learning Meets Privacy: A Survey and Outlook». Em: *ACM Computing Surveys* 54.2 (2021). DOI: 10.1145/3436755. URL: https://doi.org/10.1145/3436755.

[31] Amine Boulemtafes, Abdelouahid Derhab e Yacine Challal. «A Review of Privacy-Preserving Techniques for Deep Learning». Em: *Neurocomputing* 384 (2020), pp. 21–45. DOI: 10.1016/j.neucom.2019.11.041. URL: https://doi.org/10.1016/j.neucom.2019.11.041.

[32] Pierangela Samarati e Latanya Sweeney. *Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Suppression*. Rel. téc. SRI-CSL-98-04. SRI International, 1998. URL: https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf.

[33] Charu C. Aggarwal e Philip S. Yu, eds. *Privacy-Preserving Data Mining: Models and Algorithms*. Vol. 34. Advances in Database Systems. Springer, 2008. DOI: 10.1007/978-0-387-70992-5. URL: https://doi.org/10.1007/978-0-387-70992-5.

[34] Úlfar Erlingsson, Vasyl Pihur e Aleksandra Korolova. «RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response». Em: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 2014, pp. 1054–1067. DOI: 10.1145/2660267.2660348. URL: https://doi.org/10.1145/2660267.2660348.

[35] Noseong Park et al. «Data Synthesis Based on Generative Adversarial Networks». Em: *Proceedings of the VLDB Endowment* 11.10 (2018), pp. 1071–1083. DOI: 10.14778/3231751.3231757. URL: https://doi.org/10.14778/3231751.3231757.

[36] Martín Abadi et al. «Deep Learning with Differential Privacy». Em: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016, pp. 308–318. DOI: 10.1145/2976749.2978318. URL: https://doi.org/10.1145/2976749.2978318.

[37] Reza Shokri e Vitaly Shmatikov. «Privacy-Preserving Deep Learning». Em: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2015, pp. 1310–1321. DOI: 10.1145/2810103.2813687. URL: https://doi.org/10.1145/2810103.2813687.

[38] Keith Bonawitz et al. «Practical Secure Aggregation for Privacy-Preserving Machine Learning». Em: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2017, pp. 1175–1191. DOI: 10.1145/3133956.3133982. URL: https://doi.org/10.1145/3133956.3133982.

[39] Stacey Truex et al. «A Hybrid Approach to Privacy-Preserving Federated Learning». Em: *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec '19)*. 2019, pp. 1–11. DOI: 10.1145/3338501.3357370. URL: https://doi.org/10.1145/3338501.3357370.

[40] Benjamin C. M. Fung et al. «Privacy-Preserving Data Publishing: A Survey of Recent Developments». Em: *ACM Computing Surveys* 42.4 (2010), pp. 1–53. DOI: 10.1145/1749603.1749605. URL: https://doi.org/10.1145/1749603.1749605.

[41] Alberto Blanco-Justicia et al. «A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning». Em: *ACM Computing Surveys* 55.8 (2023), pp. 1–16. DOI: 10.1145/3547139. URL: https://doi.org/10.1145/3547139.

[42] Ashwin Machanavajjhala et al. «l-Diversity: Privacy Beyond k-Anonymity». Em: *ACM Transactions on Knowledge Discovery from Data* 1.1 (2007), p. 3. DOI: 10.1145/1217299.1217302. URL: https://doi.org/10.1145/1217299.1217302.

[43] Ninghui Li, Tiancheng Li e Suresh Venkatasubramanian. «t-Closeness: Privacy Beyond k-Anonymity and l-Diversity». Em: *Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE)*. 2007, pp. 106–115. DOI: 10.1109/ICDE.2007.367856. URL: https://doi.org/10.1109/ICDE.2007.367856.

[44] Charu C. Aggarwal. «On k-Anonymity and the Curse of Dimensionality». Em: *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*. VLDB Endowment, 2005, pp. 901–909. DOI: 10.5555/1083592.1083696. URL: https://research.ibm.com/publications/on-k-anonymity-and-the-curse-of-dimensionality.

[45] Shiva Prasad Kasiviswanathan et al. «What Can We Learn Privately?» Em: *SIAM Journal on Computing* 40.3 (2011), pp. 793–826. DOI: 10.1137/090756090. URL: https://doi.org/10.1137/090756090.

[46] Stanley L. Warner. «Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias». Em: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69. DOI: 10.1080/01621459.1965.10480775. URL: https://doi.org/10.1080/01621459.1965.10480775.

[47] John C. Duchi, Michael I. Jordan e Martin J. Wainwright. «Local Privacy and Statistical Minimax Rates». Em: *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2013, pp. 429–438. DOI: 10.1109/FOCS.2013.53. URL: https://doi.org/10.1109/FOCS.2013.53.

[48] Rakesh Agrawal e Ramakrishnan Srikant. «Privacy-Preserving Data Mining». Em: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. 2000, pp. 439–450. DOI: 10.1145/335191.335438. URL: https://dl.acm.org/doi/10.1145/335191.335438.

[49] Stephen E. Fienberg e Julie McIntyre. «Data Swapping: Variations on a Theme by Dalenius and Reiss». Em: *Privacy in Statistical Databases (PSD 2004)*. Vol. 3050. Lecture Notes in Computer Science. Springer, 2004, pp. 14–29. DOI: 10.1007/978-3-540-25955-8_2. URL: https://link.springer.com/chapter/10.1007/978-3-540-25955-8_2.

[50] Michael Hay et al. «Boosting the Accuracy of Differentially Private Histograms Through Consistency». Em: *Proceedings of the VLDB Endowment (PVLDB)*. Vol. 3. 1–2. 2010, pp. 1021–1032. DOI: 10.14778/1920841.1920970. URL: https://doi.org/10.14778/1920841.1920970.

[51] Cynthia Dwork et al. «Calibrating Noise to Sensitivity in Private Data Analysis». Em: *Theory of Cryptography Conference (TCC)*. 2006, pp. 265–284. DOI: 10.1007/11681878_14. URL: https://doi.org/10.1007/11681878_14.

[52] Claire McKay Bowen e Joshua Snoke. «Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge». Em: *Journal of Privacy and Confidentiality* 11.1 (2021). DOI: 10.29012/jpc.748. URL: https://arxiv.org/abs/1911.12704.

[53] Theresa Stadler, Bogdan Oprisanu e Carmela Troncoso. «Synthetic Data—Anonymisation Groundhog Day». Em: *31st USENIX Security Symposium (USENIX Security 22)*. 2022. URL: https://www.usenix.org/system/files/sec22-stadler.pdf.

[54] Lei Xu et al. «Modeling Tabular data using Conditional GAN». Em: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019, pp. 7333–7343. URL: https://arxiv.org/abs/1907.00503.

[55] Liyang Xie et al. «Differentially Private Generative Adversarial Network». Em: (2018). URL: https://arxiv.org/abs/1802.06739.

[56] Cynthia Dwork e Aaron Roth. «The Algorithmic Foundations of Differential Privacy». Em: *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407. DOI: 10.1561/0400000042.

[57] Arvind Narayanan e Vitaly Shmatikov. «Robust De-anonymization of Large Sparse Datasets». Em: *Proceedings of the 2008 IEEE Symposium on Security and Privacy (S&P)*. 2008, pp. 111–125. DOI: 10.1109/SP.2008.33. URL: https://doi.org/10.1109/SP.2008.33.

[58] Robin C. Geyer, Tassilo Klein e Moin Nabi. *Differentially Private Federated Learning: A Client Level Perspective*. arXiv preprint. Also presented at the NeurIPS 2017 Workshop on Private Multi-Party Machine Learning. 2017. URL: https://arxiv.org/abs/1712.07557.

[59] Nicolas Papernot et al. *Scalable Private Learning with PATE*. arXiv preprint. ICLR 2018. 2018. URL: https://arxiv.org/abs/1802.08908.

[60] Alekh Agarwal et al. «A Reductions Approach to Fair Classification». Em: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018, pp. 60–69. URL: https://proceedings.mlr.press/v80/agarwal18a.html.

[61] Sarah Bird et al. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Rel. téc. MSR-TR-2020-32. Microsoft, 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[62] Hsiang Chang e Reza Shokri. «On the Privacy Risks of Algorithmic Fairness». Em: *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. 2021, pp. 292–303. DOI: 10.1109/EuroSP51992.2021.00028. URL: https://ieeexplore.ieee.org/document/9581256.

[63] Chang Sun, Johan van Soest e Michel Dumontier. «Improving Correlation Capture in Generating Imbalanced Data using Differentially Private Conditional GANs». Em: *arXiv preprint arXiv:2206.13787* (2022). URL: https://arxiv.org/abs/2206.13787.

[64] H. Brendan McMahan et al. «Communication-Efficient Learning of Deep Networks from Decentralized Data». Em: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2017, pp. 1273–1282. URL: https://arxiv.org/abs/1602.05629.

[65] Ronald Cramer, Ivan Damgård e Jesper Buus Nielsen. *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, 2015. ISBN: 9781107043053. URL: https://www.universiteitleiden.nl/en/research/research-output/science/secure-multiparty-computation-and-secret-sharing.

[66] Yehuda Lindell. «Secure Multiparty Computation». Em: *Communications of the ACM* 64.1 (2021), pp. 86–96. DOI: 10.1145/3387108. URL: https://doi.org/10.1145/3387108.

[67] Craig Gentry. «Fully Homomorphic Encryption Using Ideal Lattices». Em: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC '09)*. 2009, pp. 169–178. DOI: 10.1145/1536414.1536440. URL: https://doi.org/10.1145/1536414.1536440.

[68] Abdulkadir Acar et al. «A Survey on Homomorphic Encryption Schemes: Theory and Implementation». Em: *ACM Computing Surveys* (2018). DOI: 10.1145/3214303.

[69] Victor Costan e Srinivas Devadas. *Intel SGX Explained*. IACR Cryptology ePrint Archive. 2016. URL: https://eprint.iacr.org/2016/086.pdf.

[70] Ahmed Salem et al. «ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models». Em: *Network and Distributed System Security Symposium (NDSS)*. 2019. DOI: 10.14722/ndss.2019.23119. URL: https://doi.org/10.14722/ndss.2019.23119.

[71] Matt Fredrikson, Somesh Jha e Thomas Ristenpart. «Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures». Em: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.

[72] David J. Hand. «Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve». Em: *Machine Learning* 77.1 (2009), pp. 103–123. DOI: 10.1007/s10994-009-5119-5.

[73] Miroslav Kubát e Stan Matwin. «Addressing the Curse of Imbalanced Training Sets: One-Sided Selection». Em: *Proceedings of the 14th International Conference on Machine Learning (ICML)*. 1997, pp. 179–186. URL: https://www.semanticscholar.org/paper/Addressing-the-Curse-of-Imbalanced-Training-Sets:-Kub%C3%A1t-Matwin/ebc3914181d76c817f0e35f788b7c4c0f80abb07.

[74] Glenn W. Brier. «Verification of Forecasts Expressed in Terms of Probability». Em: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[75] V. V. L. Divakar Allavarpu, Vankamamidi Srinivasa Naresh e A. Krishna Mohan. «Privacy-preserving credit risk analysis based on homomorphic encryption aware logistic regression in the cloud». Em: *Security and Privacy* 7.3 (2024). DOI: 10.1002/spy2.372. URL: https://doi.org/10.1002/spy2.372.

[76] Lorenzo Andolfo et al. «Privacy-Preserving Credit Scoring via Functional Encryption». Em: *Computational Science and Its Applications – ICCSA 2021*. Vol. 12956. Lecture Notes in Computer Science. Springer, 2021, pp. 31–43. DOI: 10.1007/978-3-030-87010-2_3. URL: https://doi.org/10.1007/978-3-030-87010-2_3.

[77] Haoran He et al. «A privacy-preserving decentralized credit scoring method based on multi-party information». Em: *Decision Support Systems* 166 (2023), p. 113910. DOI: 10.1016/j.dss.2022.113910. URL: https://doi.org/10.1016/j.dss.2022.113910.

[78] Vittorio Prodomo et al. «Privacy-Preserving Data Obfuscation for Credit Scoring». Em: *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC 2025)*. Association for Computing Machinery, 2025, pp. 114–121. DOI: 10.1145/3672608.3707864. URL: https://doi.org/10.1145/3672608.3707864.

[79] Oshan Mudannayake et al. «On Privacy-Preserved Machine Learning Using Secure Multi-Party Computing: Techniques and Trends». Em: *Computers, Materials & Continua* 85.2 (2025), pp. 2527–2578. DOI: 10.32604/cmc.2025.068875. URL: https://doi.org/10.32604/cmc.2025.068875.

[80] Maryam Aliakbarpour et al. «Enhancing Feature-Specific Data Protection via Bayesian Coordinate Differential Privacy». Em: *arXiv preprint arXiv:2410.18404* (2024). URL: https://arxiv.org/abs/2410.18404.

[81] Iain Brown e Christophe Mues. «An experimental comparison of classification algorithms for imbalanced credit scoring data sets». Em: *Expert Systems with Applications* 39.3 (2012), pp. 3446–3453. DOI: 10.1016/j.eswa.2011.09.033. URL: https://doi.org/10.1016/j.eswa.2011.09.033.

[82] Tianhao Wang et al. «Locally Differentially Private Protocols for Frequency Estimation». Em: *26th USENIX Security Symposium (USENIX Security 17)*. 2017, pp. 729–745. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

[83] Mei Ling Fang, Devendra Singh Dhami e Kristian Kersting. «DP-CTGAN: Differentially Private Medical Data Generation Using CTGANs». Em: *Artificial Intelligence in Medicine (AIME 2022)*. Vol. 13263. Lecture Notes in Computer Science. Springer, 2022, pp. 178–188. DOI: 10.1007/978-3-031-09342-5_17. URL: https://link.springer.com/chapter/10.1007/978-3-031-09342-5_17.