Privacy Metrics for Machine Learning Datasets: Comprehensive Evaluation and Proposal

1. Introduction

In this project, a comprehensive evaluation of privacy-preserving techniques for machine learning datasets was performed. The key goal was to implement various existing privacy metrics and combine them into a normalized weighted average metric — a new proposal for assessing dataset privacy.

The following privacy-preserving methods were explored:

- Differentially Private Stochastic Gradient Descent (DP-SGD)
- Reconstruction Attack evaluation
- Traditional privacy models: K-Anonymity, L-Diversity, and T-Closeness

2. Methodology Overview

2.1 Differential Privacy using DP-SGD

- **Library**: Opacus
- **Datasets**:
    - Credit Customers Dataset
    - Diabetes Dataset
    - Employee Dataset
- **Privacy Parameters**:
    - Noise Multiplier: 1.3–1.5
    - Max Gradient Norm: 1.0–1.2
    - δ = 1e-5
    - ε values reported for each experiment

2.2 Reconstruction Attack

- **Attack Goal**: Reconstruct original training data from model parameters
- **Metrics used**:
    - Mean Squared Error (MSE) (lower = better reconstruction)
    - Cosine Similarity (closer to 1 = better reconstruction)

2.3 K-Anonymity, L-Diversity, T-Closeness

- **Datasets**:
  - Diabetes Dataset
  - Credit Customers Dataset (for K and L)
- **Methods**:
  - Generalization via binning
  - Stratified grouping based on quasi-identifiers
  - Iterative search for minimal binning satisfying all three constraints

3. Results and Analysis

3.1 Differential Privacy - DP-SGD Results

Dataset	Model Type	Final Test Accuracy	Final ε	Privacy Notes
Credit Customers (Raw)	DP	62.26%	3.04	Good privacy (ε < 10)
	Non-DP	68.5%	N/A	Baseline accuracy
Credit Customers (ACT-GAN)	DP	65.67%	3.04	Good privacy
	Non-DP	67.5%	N/A	Slightly higher than DP
Diabetes (Raw)	DP	51.51%	4.45	Acceptable privacy
	Non-DP	72.08%	N/A	Stronger baseline accuracy
Diabetes (ACT-GAN)	DP	66.67%	4.45	Acceptable
	Non-DP	78.57%	N/A	High baseline

3.2 Reconstruction Attack Results

Dataset	Model Type	MSE	Cosine Similarity	Key Observations
Credit Customers (Raw)	DP	306.3831	0.1050	Poor reconstruction, strong privacy
	Non-DP	231.5900	0.1683	More information leakage
Credit Customers (ACT-GAN)	DP	464.8475	-0.0135	Even stronger privacy
	Non-DP	463.4412	0.0145	Minimal difference
Diabetes (Raw)	DP	1144.1663	0.0370	Extremely high MSE, strong privacy
	Non-DP	3.0803	0.2059	High leakage, risk of reconstruction
Diabetes (ACT-GAN)	DP	679.1671	0.1560	Good privacy
	Non-DP	2.4822	0.1468	Significant leakage

Interpretation:
DP models show strong resistance to reconstruction, validating the effectiveness of DP-SGD. Non-DP models are vulnerable to data leakage.

3.3 K-Anonymity, L-Diversity, and T-Closeness Results

Dataset	Privacy Models Applied	Result Summary
Diabetes (Raw)	K=5 Anonymity	Achieved with Age and Pregnancies binning
	K=5, L=2 Diversity	Successfully achieved
	K=5, L=2, T=0.2 Closeness	Successfully achieved with bin [0-17] for Pregnancies
Credit Customers	K=5 Anonymity + L=2 Diversity	Achieved with suitable age and credit amount generalizations

Interpretation:
Proper binning strategies can achieve strong traditional privacy guarantees. However, they often involve **loss of granularity** and **information utility trade-offs**.

4. Conclusion

- Differential Privacy mechanisms (like DP-SGD) provide strong formal privacy guarantees while maintaining acceptable accuracy.
- Traditional privacy models (K-Anonymity, L-Diversity, T-Closeness) remain effective but can lead to major data utility loss.
- Reconstruction attacks reveal that non-private models can significantly leak sensitive information.