

Experiment / Artefact Demonstration Document

Research Project Title:

A Comprehensive Evaluation Approach for Privacy-Preserving Data Publishing Techniques Using Composite Scoring and Decision Support

Student Information

Name: Md. Ariful Islam

ID: 2431309

Programme: MSc Cybersecurity and Digital Forensics

University: University of Wales Trinity Saint David

1. Introduction

This document provides instructions for demonstrating the software artefact developed for the research project titled “*A Comprehensive Evaluation Approach for Privacy-Preserving Data Publishing Techniques Using Composite Scoring and Decision Support.*”

The artefact implements a complete evaluation method for comparing privacy-preserving data publishing techniques using privacy, utility, and computational performance metrics. The framework generates composite scores and rankings to support the selection of the most appropriate privacy-preserving technique.

2. Project Artefact

The developed artefact consists of a Python-based experimental approach implemented in Google Colab and Jupyter Notebook environments.

The framework performs the following tasks:

- Data preprocessing and cleaning
- Feature engineering
- Privacy-preserving data transformation
- Machine learning utility evaluation
- Privacy assessment
- Computational performance analysis
- Composite score calculation
- Decision-support ranking

3. Source Code Repository

Project Code

Google Drive Repository:

Project_Code / CodeBase

<https://drive.google.com/file/d/16b2gsGd8b3CiQiUXRowUY-IKPz1Lufs3/view?usp=sharing>

4. Dataset Information

Adult Income Dataset

Reference

Becker, B., & Kohavi, R. (1996).

Adult Dataset.

UCI Machine Learning Repository.

DOI:

<https://doi.org/10.24432/C5XW20>

Dataset Location

Google Drive Dataset Repository:

https://drive.google.com/drive/folders/1VBPIXa_iph4sO1o4gdds4rGHrm00W_og?usp=sharing

adult_clean.csv:

<https://drive.google.com/file/d/1WQWxyxAIEpJezMsWVk1uREV7TMTyHnuZ/view?usp=sharing>

5. Privacy-Preserving Techniques Implemented

The framework evaluates three privacy-preserving techniques:

1. k-Anonymity

- Generalisation and binning
- Protection against record linkage attacks

2. Differential Privacy

- Laplace noise injection
- Statistical privacy protection

3. Synthetic Data Generation

- Bootstrap-based synthetic record generation
- Privacy preservation through data synthesis

6. Machine Learning Evaluation

The framework evaluates utility using classification models.

The following algorithms were compared:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

Experimental results showed that **XGBoost** achieved the highest predictive performance and was selected as the final utility evaluation model.

Performance achieved:

- Accuracy: 86.44%
- F1-Score: 70.49%

7. Composite Scoring Framework

The final ranking is generated using the following composite scoring model:

$$S = 0.4P + 0.4U + 0.2C$$

Where:

- P = Privacy Score
- U = Utility Score
- C = Computational Performance Score

The weighting scheme gives equal importance to privacy and utility while considering computational efficiency as a supporting factor.

8. Demonstration Procedure

To reproduce the experiment:

Step 1

Download the project code from the provided repository.

Step 2

Download the Adult Income Dataset from the dataset repository.

Step 3

Open the notebook in Google Colab or Jupyter Notebook.

Step 4

Install required Python libraries:

- pandas
- numpy
- scikit-learn
- xgboost
- matplotlib
- seaborn
- psutil

Step 5

Execute notebook cells sequentially.

Step 6

Review generated outputs including:

- Utility evaluation metrics
- Privacy scores
- Computational performance metrics
- Composite scores
- Final ranking of privacy-preserving techniques

9. Expected Results

The experimental evaluation identified:

1. Synthetic Data Generation as the highest-ranked technique.
2. k-Anonymity as the second-ranked technique.
3. Differential Privacy as the third-ranked technique.

Under the final 80:20 train-test split configuration, Synthetic Data Generation achieved the highest composite score and provided the most balanced trade-off between privacy, utility, and computational performance.

10. Conclusion

The developed artefact demonstrates a practical decision-support framework for evaluating privacy-preserving data publishing techniques. By integrating privacy, utility, and computational performance into a unified composite scoring model, the framework enables systematic comparison and ranking of privacy-preserving methods for secure data sharing and machine learning applications.