IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 9, May 2025

Impact Factor: 7.67

Data Comparison Tool

Ayush Mishra, Ansh Tyagi, Siddharth Jha, Abhishek Kumar
Department of CSE and ML
Raj Kumar Goel Institute of Technology, Ghaziabad, India

Abstract: *In recent years, synthetic data has become increasingly popular due to its wide range of applications. However, evaluating the usefulness and privacy of synthetic data remains a complex task, as existing assessment methods vary based on the type of data, the intended use case, and the evaluation goal. Currently, there is no universal method for measuring utility and privacy together. To address this, we have created a comprehensive tool that combines 24 modern evaluation techniques into a single executable program. This tool generates reports that compare datasets—synthetic or real—by highlighting similarities and identifying potential privacy risks. The included methods span from visual and statistical utility assessments to privacy risk detection, making it easier for researchers and industry professionals to understand and improve their synthetic datasets. Our goal is to offer a solid foundation for developing more general and reliable evaluation methods in the future.*

Keywords: Synthetic data, data utility, data privacy, dataset comparison, evaluation metrics

I. INTRODUCTION

Synthetic data refers to information that mimics real data but isn't directly generated from actual events or processes. Instead, it is created through algorithms or models that simulate real-world characteristics. The advantage is that it can be used in similar ways to real data without violating privacy laws or needing actual personal data.

Although synthetic data is still underutilized in healthcare, its applications are growing. It's currently being used in areas such as:

- **Software testing** – to simulate test cases during development.
- **Training and education** – for onboarding new employees or training medical students.
- **Machine learning** – for tasks like data augmentation or algorithm testing.
- **Regulatory compliance** – especially as AI becomes more common in healthcare.
- **Data retention** – replacing sensitive data with synthetic versions after deletion.
- **Academic and industrial data sharing** – enabling safe data access for research.

A common promise of synthetic data is better privacy protection. However, improving privacy usually means reducing the data's usefulness. Therefore, finding a balance between utility and privacy remains a challenge. Measuring either of them still lacks a universal standard and depends heavily on the context.

II. METHODS

We developed a Python-based data analysis tool that compares two datasets and produces a detailed PDF report. It uses libraries such as SciPy, scikit-learn, sdmetrics, and mlxtend. This tool can also be turned into an API for integration into other systems.

Key Steps in the Pipeline:

- **Dataset loading and cleanup** – including removal of columns with too many missing values.
- **Visual utility analysis** – using bar charts, density plots, heatmaps, and pair plots.
- **Quantitative utility analysis** – includes statistical tests and distance metrics:
    - *Column-wise*: Kolmogorov–Smirnov test (for continuous), Chi-square test (for categorical), entropy, and various divergence/distance metrics.
    - *Pair-wise*: Kullback–Leibler divergence between column pairs.

DOI: 10.48175/568

    - *Table-wise*: Log-likelihoods from Gaussian Mixture Models or Bayesian Networks, plus machine learning model performance comparisons.
- **Privacy risk assessment** – includes duplicate detection, record linkage, and distance measures (cosine, Euclidean, matrix).

The pipeline uses the following libraries and tools:

- **SciPy** for statistical testing and distribution analysis
- **Scikit-learn** for machine learning models and preprocessing
- **SDMetrics** for synthetic data-specific evaluations
- **MLxtend** for helper functions such as model comparison and data transformation
- **LaTeX** to compile a professional-quality PDF report

Machine Learning-Based Utility Testing

To explore model-level similarities, we:

- Train linear regression and decision tree models on both datasets.
- Compare performance metrics (like RMSE, accuracy, or R²) to evaluate consistency.
- Use **cross-validation across datasets**—train on real, test on synthetic, and vice versa—to assess generalizability.

Metric Summary Table

Each metric is categorized based on its purpose—utility or privacy—and whether it's visual or quantitative. This organized approach helps users navigate the report and understand how each method contributes to the overall evaluation

III. RESULTS

We tested our tool using a publicly available heart disease dataset from the UCI repository. Synthetic data was generated using the synthpop R package. Our tool produced visual plots and statistical summaries to compare the real and synthetic datasets across various utility and privacy metrics. The full source code and examples are available on GitHub.

To evaluate the effectiveness and coverage of our dataset comparison tool, we applied it to a well-known public dataset: the **Heart Disease dataset from the UCI Machine Learning Repository**. Using this real dataset as a baseline, we generated synthetic data using the synthpop package in R. The goal was to compare both datasets across utility and privacy dimensions using the full suite of evaluation techniques built into our pipeline.

IV. DISCUSSION AND CONCLUSION

The methods we compiled can help compare datasets beyond just real vs. synthetic. For instance, they are valuable in federated learning environments, where data is spread across multiple locations (silos). By evaluating column-level and inter-column similarities, one can better understand population differences and similarities across institutions.

Our goal was to provide an accessible and consistent benchmark to evaluate dataset similarity and help foster safer, more effective use of synthetic data in healthcare and beyond. Ultimately, we hope this leads to the development of standardized, widely accepted evaluation methods.

This work represents an important step toward a **standardized framework** for evaluating synthetic data. By unifying a diverse set of metrics under a single tool and making the results accessible through automated reports, we aim to **empower researchers, data scientists, and policy makers** to make more informed decisions about synthetic data usage. Ultimately, our goal is to contribute to the development of a **broader, standardized methodology** for dataset comparison and to promote **ethical and effective use** of synthetic data in sensitive domains like healthcare.

In conclusion, our work presents a comprehensive and flexible tool for evaluating the utility and privacy of synthetic datasets, addressing a critical gap in current data science practices. By integrating a wide range of qualitative

DOI: 10.48175/568

visualizations and quantitative statistical tests, our solution provides a holistic assessment of how closely synthetic data resembles real datasets, both in structure and behavior. The tool's modular design, built on widely-used Python libraries and supporting automated PDF reporting, makes it suitable for diverse applications in healthcare, machine learning, regulatory compliance, and more. Our case study using the UCI Heart Disease dataset demonstrates the practical value of the tool in identifying similarities and potential privacy risks. Furthermore, the tool can be adapted to various use cases, helping institutions ensure data quality while maintaining privacy. This is especially important in domains like healthcare, where data sensitivity is paramount. While the current version supports structured data well, future work will focus on expanding capabilities to handle time-series and unstructured data, as well as enhancing automation features

DOI: 10.48175/568