Generalised Weighted Framework for Synthetic Data Evaluation [1]

Chiung Ching Ho BNM

1 This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, Bank Negara Malaysia, the BIS, the IFC or the other central banks and institutions represented at the event.

Overview

Privacy, utility and fidelity for trust-worthy synthetic data

There is a need to use synthetic data by Central Banks
driven by both legislation and a need to maintain privacy of data

Synthetic data produced by Central Banks needs to be trusted
to fulfil privacy, utility and fidelity (PUF) requirements

An assessment framework is proposed to evaluate synthetic data
produced by synthetic data generators (SDG) that assesses PUF requirements using a flexible and extensible framework

Index

1. Introduction
   - ❑ Research motivation
   - ❑ Synthetic data sharing and central banks
2. Considerations for synthetic data
   - ❑ Privacy, utility , fidelity
3. Techniques
4. Framework for synthetic data evaluation
5. Experiment
   - ❑ Data
   - ❑ Results
6. Discussion
7. Summary and recommendations
8. References
9. Addendum

Generalised Weighted Framework for Synthetic Data Evaluation

Motivation : Inspired by an attempt to create a universal metric for evaluation of synthetic data [1], a new framework is proposed that will allow for flexible assessment of synthetic data from a privacy, utility and fidelity (P.U.F) perspective

Problem statement : Data created by SDG needs to be evaluated for privacy, utility and fidelity (PUF) to ensure that synthetic data is credible to be shared or published by central banks or authorities, and current solutions does not allow for flexibility in measuring SDG that balances P.U.F needs

Research Question (RQ) 1 : Which of current SDG techniques exhibit good PUF scores?

Research Question (RQ) 2 : How do we balance P.U.F measures for different SDG to address different analytical needs?

Assessing synthetic data

- Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data-science task(s)[2]
- Synthetic data is being used by central banks to enable data sharing without compromising on privacy or infringing on legislation[3]
- Reasons for sharing data using synthetic data in central banks includes
  - Sharing micro data for research purposes[4]
  - Justification of data collection rigor [5]
  - Data disclosure risk mitigation and[6]
  - Sharing of granular data instead of aggregated data[7]
- Examples of synthetic data shared by central banks includes micro data from surveys, non-financial firm-level data, and loans to legal persons[4][5][6][7]

Synthetic data generation (SDG) techniques comparison

Name of Technique	Strengths	Weaknesses
Gaussian Diffusion Models (GDM)	High flexibility in modelling distributions Good for high-dimensional data	Computationally intensive Requires large amounts of training data
Gaussian Copula (GC)	Maintains statistical properties Suitable for tabular data	Assumes a Gaussian dependence structure Struggles with complex, non-linear relationships
Conditional Tabular GAN (CTGAN)	Handles imbalanced and multi-modal data Captures complex dependencies	Training instability Sensitive to hyperparameter tuning
Tabular Variational Autoencoder (TVAE)	Effective in capturing latent structures Suitable for tabular data	May overfit on small datasets Needs careful parameter tuning
Gaussian Mixtures Models (GMM)	Models data as a combination of distributions Interpretable and straightforward	Assumes data can be represented as Gaussian mixtures Sensitive to initialization
Time Dependent Self Attention (TDSA)	Strong at time-series data generation Preserves temporal correlations	Limited to time-series data Computationally expensive for large datasets
Tabular Diffusion Probabilistic Model (TabDPM)	Robust for tabular data with complex dependencies Flexible with varying distributions	Computationally heavy Requires careful parameter configuration

Table 1 : A comparison of SDG techniques

Assessing synthetic data : P.U.F for credible synthetic data

Privacy	Utility	Fidelity
❑ Is x in the dataset? ❑ Can unknown data about x be found? ❑ Can I recreate the original data?	❑ How does the synthetic data perform as compared to the original data in specific task?	❑ How truthful is the synthetic data? ❑ How statistically similar is the synthetic data? ❑ How similar is the distribution of the synthetic data?

There is always a trade-off between privacy and utility, while fidelity may proxy for utility

Generalised Weighted Framework for Synthetic Data Evaluation

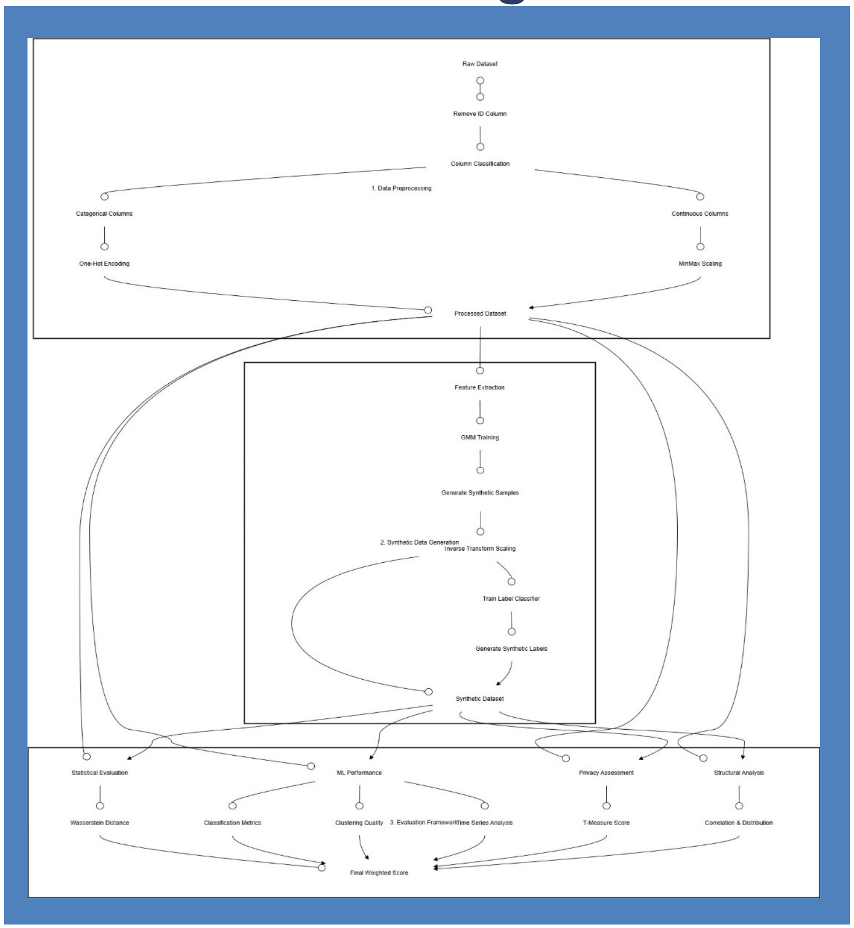


Figure 1: Synthetic data evaluation process flow

Phase 1 - Data Preprocessing
•Start with the raw dataset
•Remove any ID columns
•Classify columns into two types:
- Categorical (like SEX, EDUCATION, MARRIAGE, PAY status)
- Continuous (like AGE, BILL amounts, PAYMENT amounts)

•Process each type differently:
- Categorical columns get one-hot encoded
- Continuous columns get scaled using MinMax scaling

•Result: A cleaned, processed dataset

1.Preprocessing is dependent on data set

Phase 2 - Synthetic Data Generation
•Take the processed dataset
•Extract features for GMM training
•Train the Gaussian Mixture Model
•Generate synthetic samples
•Convert the synthetic data back to original scale
•Generate appropriate labels using a classifier
•Result: A synthetic dataset that mimics the original

2. The synthetic data generator (SDG) is chosen based on use case

Phase 3 - Evaluation
•Compare real and synthetic data across four aspects:
- Statistical similarity (using Wasserstein distance)
- Machine learning performance (classification accuracy, clustering quality, time series analysis)
- Privacy assessment (using T-measure score)
- Structural analysis (correlations and distributions)

•Combine all metrics into a final weighted score
Each metric contributes equally (25%) to the final evaluation score, which helps assess how well the synthetic data captures the important characteristics of the original dataset while maintaining privacy.

3. Privacy, Utility (ML scores) and Fidelity (statistical and structural) measures can be changed or weighted as needed

Figure 2: Synthetic data evaluation measures using Gaussian Mixture Models as a Synthetic Data Generator

Experiment using credit card data

The synthetic data was created from an open source credit card dataset*
This dataset was chosen as it was suitable for assessing utility through machine learning applications
It has 30,000 entries (excluding header) containing information about credit card clients. Here are the key features:

Financial

- LIMIT_BAL: Credit limit balance
- BILL_AMT1 through BILL_AMT6: Bill amounts for six consecutive months
- PAY_AMT1 through PAY_AMT6: Payment amounts for six consecutive months

Demographics

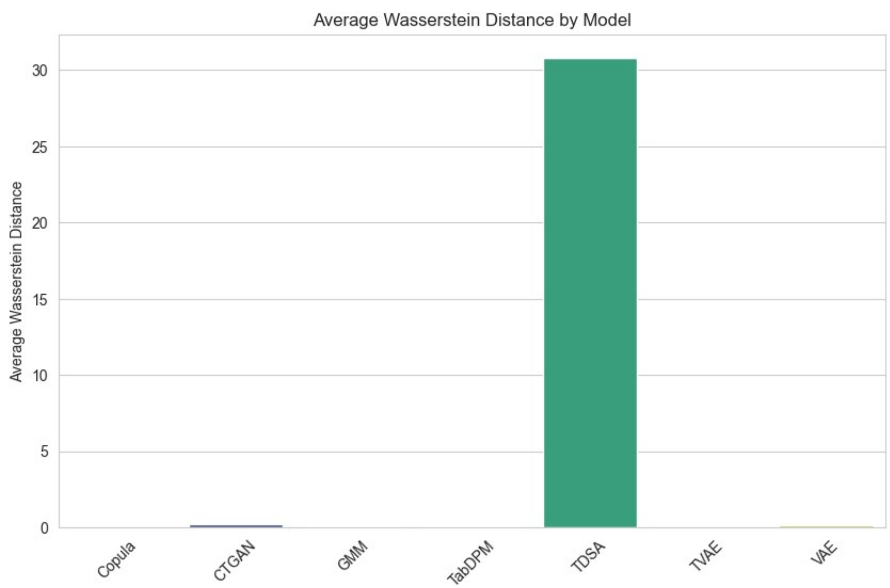
- SEX: Gender of the client
- AGE: Age of the client
- EDUCATION: Education level
- MARRIAGE: Marital status

Historical payment

- PAY_0 through PAY_6: Payment status history for seven months
- "Default payment next month": Binary indicator (0 or 1) predicting whether the client will default on their payment in the next month

* Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2, Part 1), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Fidelity : Measuring Wasserstein distance and data correlation & distribution



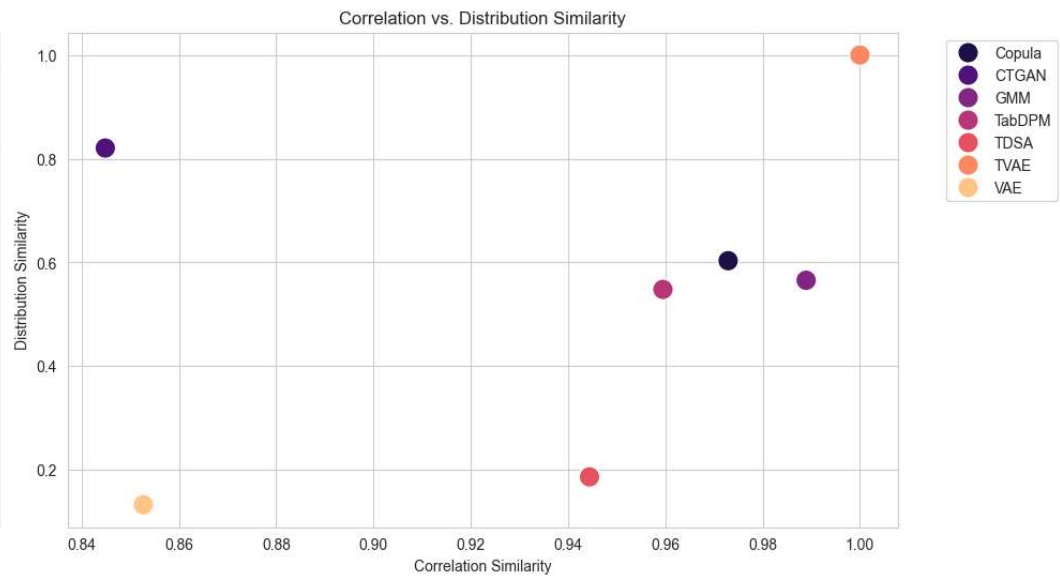


Figure 3 : Average Wasserstein distance for all columns for synthetic data relative to original data except for TDSA

Figure 4 : TVAE generated synthetic data has high correlation and distribution similarity

Utility : Classification and Clustering and Time Series Forecast

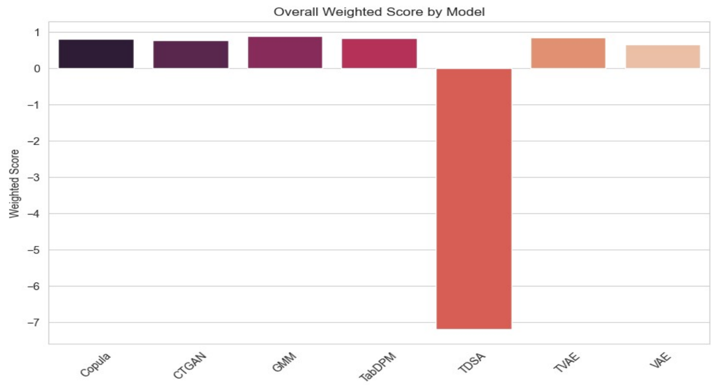


Figure 5 : Overall weighted scores for synthetic data produced by SDGs are similar except for TDSA

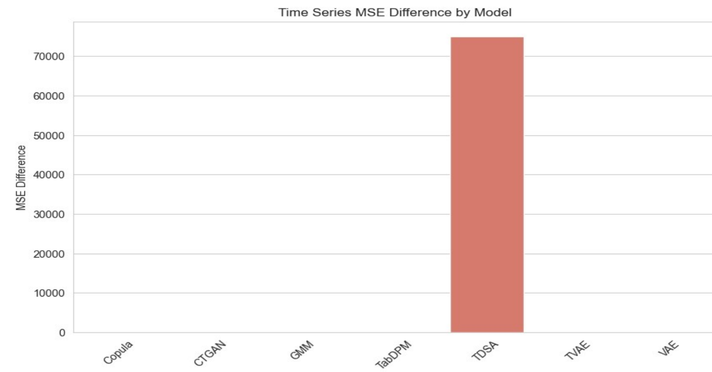


Figure 6: Time series forecast mean square error for synthetic data produced by various SDG with TSDA as an outlier

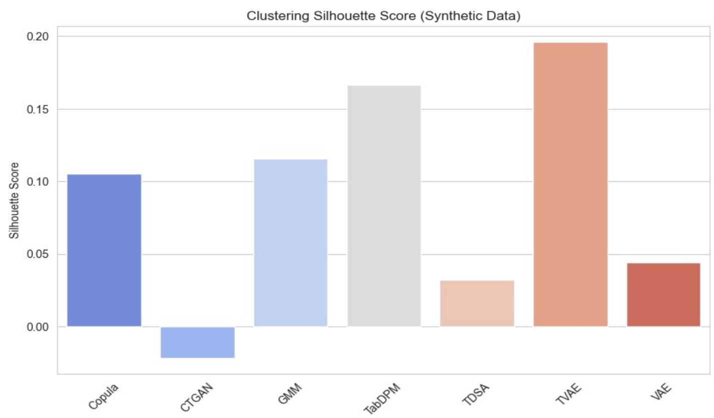


Figure 7: Clustering scores for synthetic data produced by various SDGs are generally poor

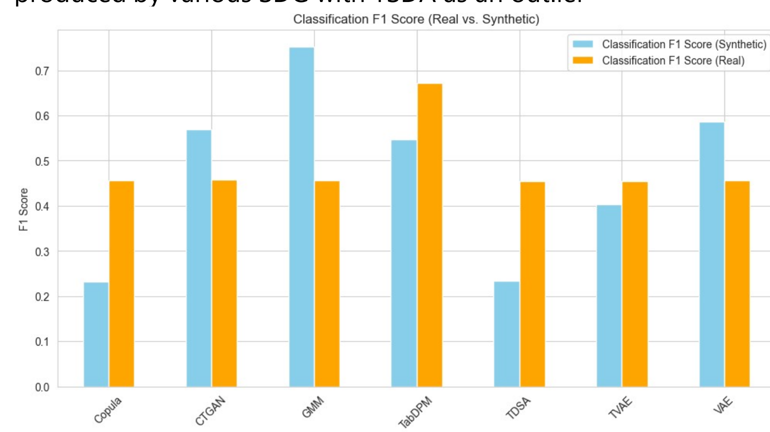


Figure 8: F1 scores for synthetic data produced by SDGs

Privacy : T-measure scores for various synthetic data produced by various SDG

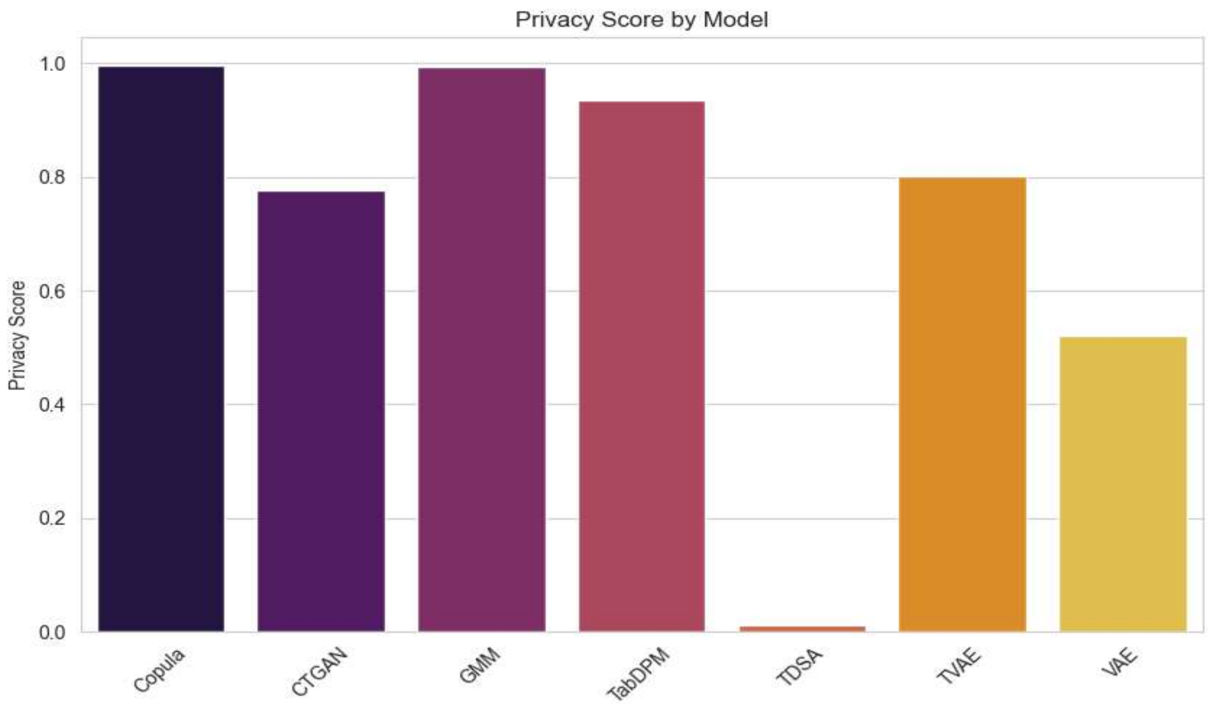


Figure 9: Privacy scores for datasets produced by various SDG with TDSA as an outlier

T-measure is the absolute difference between the synthetic and original column mean for all numeric columns, normalised by the standard deviation. This is then transformed into the privacy score

Privacy vs Utility trade-off : Choosing the SDG that fulfils both

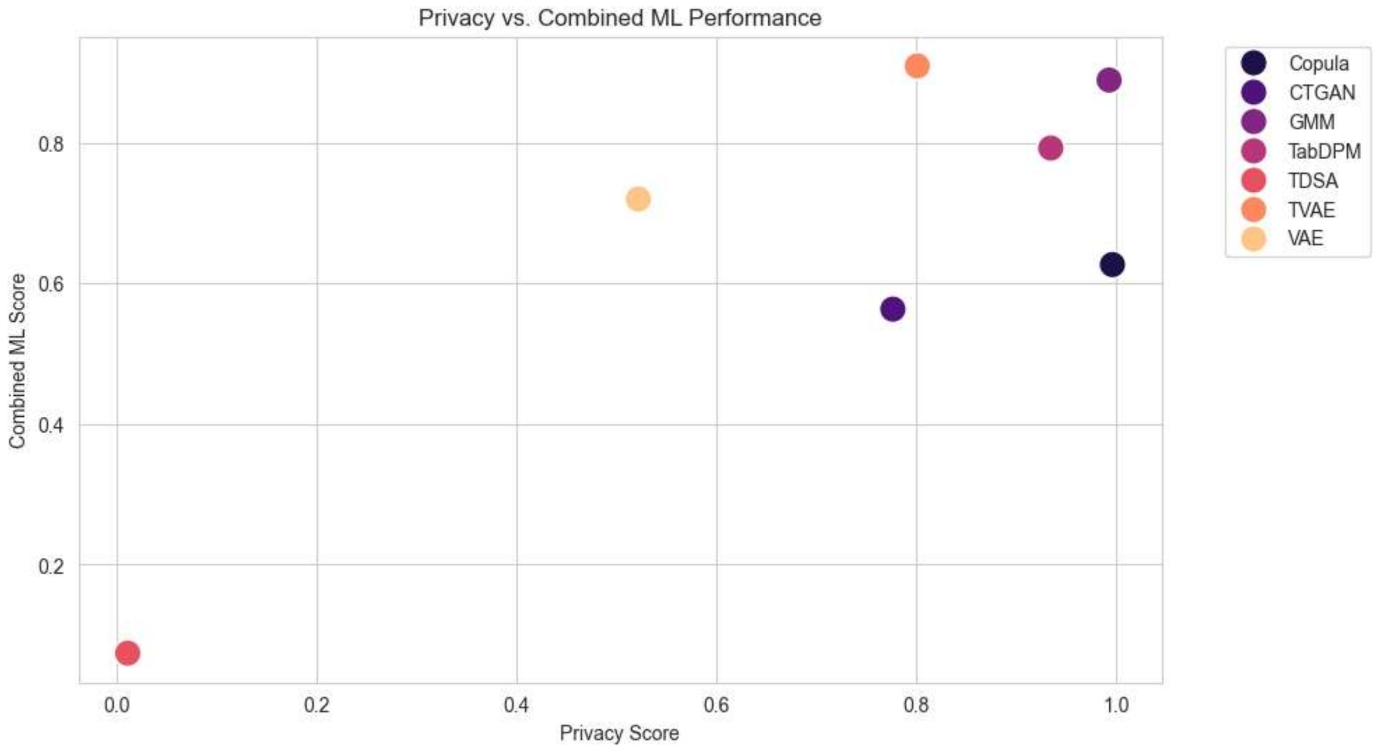


Figure 10: Privacy vs Utility (Combined ML) Score plot

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

Discussion : Weighted score aligns with privacy, utility and fidelity measures

Model	Weighted Score	Privacy	Utility	Fidelity	Runtime(s)	Evaluation Summary
GMM	0.89	High	High	Moderate	34.5	Best overall performer. High privacy and utility and moderate fidelity. Highest weighted score.
TVAE	0.86	High	High	High	1113.2	High privacy, high utility, and high fidelity. Weighted score reflects this balance.
TabDPM	0.83	High	Moderate	Moderate	365	High privacy with moderate utility and fidelity. Weighted score reflects this balance.
Copula	0.82	High	Moderate	Moderate	84448.3	High privacy, but moderate utility and fidelity. Weighted score reflects this balance.
CTGAN	0.77	Moderate	Low	High	139.3	Moderate privacy, low utility but high fidelity. Weighted score aligns with this.
VAE	0.65	Low	Moderate	Low	97.5	Low privacy, moderate utility and low fidelity. Weighted score aligns with this.
TDSA	-7.19	Low	Low	Low	2367.6	Poor in all dimensions. Negative weighted score reflects its poor performance.

Table 2 : Weighted scores and PUF scores alignment

Conclusion

Summary

- The proposed framework is generalisable to support different measures for different analytics tasks, answering RQ 1 & 2.
- Preprocessing of data is however difficult to generalise as it is very task-dependent.
- Future expansion of the framework may include support for non-structured data, federated machine learning and addition of differential privacy mechanism

Recommendations

1. **Privacy-Critical Applications**:
   - ❑ **GMM**: Best for high privacy, fidelity, and minority class preservation. Validate to avoid overfitting
   - ❑ **Copula**: Use for analysis in sensitive domains; at a cost of long processing times
2. **Utility-Focused Applications**:
   - ❑ **GMM**: Best for high privacy, fidelity, and minority class preservation. Fast processing times
   - ❑ **TVAE**: Strong for utility and privacy in sensitive domains
3. **Avoid**:
   - **TDSA**: Poor in all metrics; not recommended for imbalanced datasets and datasets with non-sequential data

References

[1] Chundawat, V. S., Tarun, A. K., Mandal, M., Lahoti, M., & Narang, P. (2024). TabSynDex: A Universal Metric for Robust Evaluation of Synthetic Tabular Data (arXiv:2207.05295). arXiv. https://doi.org/10.48550/arXiv.2207.05295

[2] European Union. (2024). Article 10: Data and Data Governance | EU Artificial Intelligence Act. https://artificialintelligenceact.eu/article/10/

[3]Jordon, J., Houssiau, F., Cherubin, G., Cohen, S. N., Szpruch, L., & Bottarelli, M. (2022). *Synthetic Data—What, why and how?* Alan Turing Institution.

[4] Bender, S., & Staab, P. (2015). *The Bundesbank's Research Data and Service Centre (RDSC)—Gateway to treasures of micro data on the German Financial System*. IFC workshop on "Combining micro and macro statistical data for financial stability analysis. Experiences, opportunities and challenges".

[5] Caceres, H. E., & Moews, B. (2024). *Evaluating utility in synthetic banking microdata applications* (arXiv:2410.22519). arXiv. https://doi.org/10.48550/arXiv.2410.22519

[6] Lapteva, E. K. (2023). *Utility and confidentiality assessment of synthetic company data*. IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks".

[7] Ong, C. A., Escalanda-Lo, C., Gabriel, M., & Reyes, R. (2024). Research for All: Exploring machine learning applications in generating synthetic datasets. *2024 Philippines Statistical Association Annual Conference*. https://events.psai.ph/2024/ac/

Addendum : Reflection on generative AI (GenAI) tools for enabling research (as of 31.12.2024)

	Research	Coding	Writing	Analysis
Helps	Generate new ideas	Speed up coding	Conciseness and tone	Quick initial comparisons
Caveats	Hallucination and recency	LLMs are throttled LLMs needs management	Loss of personality	Surface level analysis
Future hopes	Integrated into citation management	More efficient models to enable self-service coding	Increased support for academic writing	Increased evidence of reasoning