**Synthetic Data Generation Open-source Python Libraries Comparison: SDV and Synthicity**

**Abstract**

The growing reliance on Large Language Models (LLMs) and other machine learning systems has made access to high-quality data essential. However, real data is often scarce or difficult to obtain, particularly for smaller organizations or early-stage startups. Synthetic data generators offer a promising alternative by attempting to replicate the statistical and structural properties of real datasets while preserving utility and privacy. This study focuses on tabular synthetic data generators, comparing the performance of six models from two widely used open-source libraries: SDV (Gaussian Copula, CTGAN, TVAE) and Synthicity (Bayesian Network, CTGAN, TVAE). A real-world dataset containing electrical energy consumption and environmental data from Belgium, publicly available through the UCI Machine Learning Repository, was used. To simulate a low-data regime, only the first 1,000 rows were used for training. Each model was evaluated in two scenarios: one in which it generated 1,000 synthetic rows (1:1 ratio with the real input), and another in which it produced 10,000 synthetic rows from the same 1,000 real samples (1:10 ratio). The generated datasets were evaluated using two criteria. First, statistical similarity was assessed column-by-column using classical statistics and distributional distance measures such as the Kolmogorov–Smirnov test and the Wasserstein distance. A custom scoring system, ranging from 0 to 100, was used for interpretability. In the 1:10 scenario, all models scored between 75 and 80, with the Bayesian Network from Synthicity performing best at 80.69. Scores were substantially higher in the 1:1 scenario, with the same model achieving 96.53. Then, the predictive utility was evaluated using a “Train on Synthetic, Test on Real” strategy. Four regressors (XGBoost, Random Forest, SVR, and Linear Regression) were trained on synthetic data and tested on real data using cross-validation. The results were summarized using a custom metric ranging from negative infinity to 1, where 1 indicates parity with models trained on real data. In the 1:10 case, TVAE from SDV performed best (0.31), while CTGAN performed poorly in both libraries. In contrast, performance in the 1:1 case improved significantly across the board, with the Bayesian Network from Synthicity nearly matching real data (0.97). These results highlight a key limitation in current generative models: while statistical similarity can be achieved even with limited data, generating large quantities of high-fidelity data remains a challenge. The dramatic performance gap between the 1:1 and 1:10 scenarios suggests that scaling synthetic data from small samples still undermines its predictive value. Overall, no major performance differences were observed between SDV and Synthicity across the tested metrics. However, it is worth noting that SDV offers significantly better documentation and ease of use, which can be a decisive factor for researchers and practitioners adopting these tools in real-world settings.

**References**

[1] OpenAI. *ChatGPT.* 2022. Retrieved from <https://openai.com/index/chatgpt/>.

[2] A. Shaji George, & A. S. Hovan George. (2023). A Review of ChatGPT AI’s Impact on Several Business Sectors. *Partners Universal International Innovation Journal*, *1*(1), 9–23. <https://doi.org/10.5281/zenodo.7644359>

[3] Chen, Hao, et al. "On the Diversity of Synthetic Data and Its Impact on Training Large Language Models." *arXiv* preprint arXiv:2410.15226v2 [cs.CL], 22 Oct. 2024.

[4] Brown, Tom, et al. “Language Models Are Few-Shot Learners.” *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901, <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967411061b3e5a3861b-Paper.pdf>.

[5] OpenAI. *GPT‑4 Technical Report.* 14 Mar. 2023, <https://cdn.openai.com/papers/GPT-4-System-Card.pdf>.

[6] Bai, Yuntao, et al. *Constitutional AI: Harmlessness from AI Feedback.* Anthropic, 2022, <https://arxiv.org/abs/2212.08073>.

[7] Bubeck, Sébastien, et al. “Sparks of Artificial General Intelligence: Early Experiments with GPT‑4.” *arXiv* preprint arXiv:2303.12712 [cs.LG], 2023.

[8] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. arXiv. preprint arXiv:2211.04325, 2022.

[9] Perełkiewicz, Michał, and Rafał Poświata. "A Review of the Challenges with Massive Web-mined Corpora Used in Large Language Models Pre-Training." arXiv preprint arXiv:2407.07630 (2024).

[10] Lisa Bauer and Mohit Bansal. Identify, align, and integrate: Matching knowledge graphs to commonsense reasoning tasks. arXiv preprint arXiv:2104.10193, 2021

[11] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. arXiv preprint arXiv:2404.07503, 2024.

[12] Lu, Yingzhou, et al. *Machine Learning for Synthetic Data Generation: A Review.* arXiv preprint, arXiv:2302.04062, 30 June 2024.

[13] Bauer, André, et al. *Comprehensive Exploration of Synthetic Data Generation: A Survey.* University of Chicago, University of Würzburg, University of Ulm, Argonne National Laboratory, 2024. arXiv preprint arXiv:2401.02524.

[14] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in Neural Information Processing Systems* (NeurIPS), vol. 27, 2014, pp. 2672–2680.

[15] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." *arXiv preprint* arXiv:1511.06434, 2015.

[16] Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint* arXiv:1411.1784, 2014.

[17] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein GAN." *arXiv preprint* arXiv:1701.07875, 2017.

[18] Kingma, Diederik P., and Max Welling. "Auto-Encoding Variational Bayes." *arXiv preprint* arXiv:1312.6114, 2013.

[19] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* (NeurIPS), vol. 33, 2020, pp. 6840–6851, arXiv:2006.11239.

[20] Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems* (NeurIPS), vol. 30, 2017, pp. 5998–6008.

[21] Akiya, Ippei, Takuma Ishihara, and Keiichi Yamamoto. "Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study." JMIR Medical Informatics, vol. 12, 2024, e55118. <https://doi.org/10.2196/55118>.

[22] Jaworski, P., Durante, F., Hardle, W. K., & Rychlik, T. (2010). Copula theory and its applications (Vol. 198). New York: Springer.

[23] Patki, Neha, et al. "The Synthetic Data Vault." *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2016. *IEEE Xplore*, doi:10.1109/DSAA.2016.49.

[24] Young, Jim, et al. "Using Bayesian Networks to Create Synthetic Data." *Journal of Official Statistics*, vol. 25, no. 4, 2009, pp. 549–567.

[25] Zhang, Jun, et al. "PrivBayes: Private Data Release via Bayesian Networks." *ACM Transactions on Database Systems*, vol. 42, no. 4, 2017, pp. 1-41.

[26] Xu, Lei, et al. "Modeling Tabular Data Using Conditional GAN." *Advances in Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019.

[27] Radford, Alec, et al. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." *arXiv preprint arXiv:1511.06434* (2015).

[28] Veeramachaneni, Kalyan, et al. "CTAB-GAN+: Enhancing Tabular Data Synthesis." *Frontiers in Big Data*, vol. 6, 2023.​

[29] Gulrajani, Ishaan, et al. "Improved Training of Wasserstein GANs." *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5767–5777.

[30] Lin, Zinan, et al. "PacGAN: The Power of Two Samples in Generative Adversarial Networks." *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 14953–14963.

[31] Yadav, Parul, et al. "Rigorous Experimental Analysis of Tabular Data Generated using TVAE and CTGAN." *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 4, 2024.