2023
Volume 1
Article 5
https://doi.org/10.59453/KHZW9006

Synthetic data generator for student data serving learning analytics: A comparative study

Chen Zhan[1], Oscar Blessed Deho[2], Xuwei Zhang[1], Srećko Joksimović[1], and Maarten de Laat[1]

[1] Centre for Change and Complexity in Learning, University of South Australia·

[2] UniSA STEM, University of South Australia, Australia

Ongoing digital transformation in the education sector has led to an increased focus on learning analytics (LA). LA collects and uses students' data to gain insights about students' learning and to guide interventions and feedback. Although LA holds tremendous promise for enhancing teaching and learning, there are persistent concerns about the privacy and ethical ramifications of collecting and using student data. One potential solution is the use of Synthetic Data Generators (SDGs) which can learn from real data to generate synthetic data that closely resembles real data. This paper examines the performance of existing SDGs with student data, as well as their capabilities for serving LA. A comparative study was conducted by applying different SDGs in Synthetic Data Vault to real-world student data. We report the efficiencies of different generators and the statistical similarities between synthetic and real data. We test how well SDGs imitate the real student data by fitting generated synthetic data into commonly-used LA models. We evaluate the utility of synthetic data by the alignment of LA outputs trained using synthetic data to the ground truth of student learning outcomes recorded in real data, as well as with outputs of LA models trained by real data.

Keywords: learning analytics, privacy and ethics, synthetic data generation

Correspondence: Chen Zhan, Chen.Zhan@unisa.edu.au

Recommended citation: Zhan, C., Deho, O. B., Zhang, X., Joksimović, S., & De Laat, M. (2023). Synthetic data generator for student data serving learning analytics: A comparative study. *Learning Letters*, *1*, Article 5. https://doi.org/10.59453/KHZW9006

Introduction

Over the past few decades, rapid advancements in technology have brought about a digital transformation in our society (Zain, 2021). The education sector is no exception to this trend. The COVID-19 pandemic has further accelerated this process as the sudden shift to remote learning has forced educators to embrace technology, leading to the adoption of Learning Analytics (LA) at an unprecedented pace (Celik, Gedrimiene, Silvola, & Muukkonen, 2022). Moreover, recent developments in Artificial Intelligence (AI), such as ChatGPT (Radford, Narasimhan, Salimans, Sutskever, et al., 2018; Radford et al., 2019; Brown et al., 2020), have had a significant impact on both teachers and learners, increasing awareness about the potential of AI to revolutionise the way we teach and learn. As a result, there is growing interest in exploring how AI-powered tools can be integrated into education to enhance the learning experience and improve student outcomes (Kavitha & Lohani, 2019; Barrett et al., 2019).

The collection and utilisation of student data lie at the heart of any LA system (Joksimovic et al., 2022). However, this emphasis on data collection inevitably raises concerns about privacy and ethical implications. Privacy concerns in LA are largely associated with the extensive collection of personal information, including academic performance, behavioral patterns, and demographic data (Mutimukwe, Viberg, Oberg, & Cerratto Pargman, 2022). Many research studies and policies have emphasised the need to address these concerns as a crucial aspect of LA development (Pardo & Siemens, 2014; Drachsler & Greller, 2016; Tsai, Whitelock-Wainwright, & Gašević, 2020). In addition, the use of student data in LA raises ethical questions regarding the potential exposure of students to untested or hypothesised conditions that may negatively impact their learning experience and outcomes (Prinsloo & Slade, 2017, 2016; Benvenuti & Mazzoni, 2020). Therefore, it is essential that the use of student data for research purposes follows established procedures for ethical approval and adheres to ethical guidelines to safeguard against any risks of harm or exploitation of students.

Hence, Synthetic Data Generators (SDGs) that learn from real data to generate synthetic data that closely matches the statistical characteristics of the original data - can be viewed as a methodological innovation that addresses privacy and ethical concerns in LA research. Synthetic student data provides strong privacy guarantees and avoids ethical debates as it does not contain actual observations of students. However, the effectiveness of synthetic data in serving LA modelling remains an obstacle to its integration (Joksimovic et al., 2022). Research in synthetic data generation has gained momentum (El Emam, Mosquera, & Hoptroff, 2020), and open-source frameworks such as Synthetic Data Vault (SDV) (Patki, Wedge, & Veeramachaneni, 2016) provide user-friendly ways to generate synthetic data. Meanwhile, there is only a handful of studies with a particular focus on extolling the advantages of synthetic data (e.g., overcoming ethical barriers and benefiting data governance) (Berg, Mol, Kismihok, & Sclater, 2016a; Dorodchi, Al-Hossami, Benedict, & Demeter, 2019) and organising various use cases related to synthetic data in LA contents (Flanagan, Majumdar, & Ogata, 2022; Berg, Mol, Kismihok, & Sclater, 2016b). Less attention was paid to the technical aspects of SDGs (Vie, Rigaux, & Minn, 2022) or the application of state-of-the-art machine learning SDG technologies to enhance LA.

In this paper, we examine existing SDGs from the broader community in terms of their performances with student data, as well as their capabilities in the LA domain. A comparative study is conducted by applying a set of different SDGs in SDV, an open-sourced synthetic data generation ecosystem of libraries, to real-world student data from an Australian university. We report the efficiencies of different generators and the qualities of generated synthetic datasets regarding their statistical properties against real data. Furthermore, we test how well SDGs can provide data utility for LA modelling by fitting generated synthetic datasets into commonly-used LA models. By aligning with the ground truth of student learning outcomes recorded in real data, we evaluate the performances of LA models trained by synthetic datasets as indicators of their utilities of serving LA models.

Methodology and experiments

As an initial study investigating the use of SDGs on student data for LA, we focused on generators suited for single-table data, which is commonly encountered in student data after feature engineering in LA. SDV offers the most extensive options of SDGs for single-table data that can handle various value types (including missing values) and allow constraints to be defined over columns to adhere to certain logical rules. The SDGs used in our study are as follows:

- GaussianCopula: GaussianCopula is a statistical model to measure the dependence structures between variables by comparing the joint distribution of the variables and their marginal distributions. It assumes variables are normally distributed after a certain transformation. Thus, GaussianCopula (Wan, Li, Guo, & Zhao, 2019) aims to understand relationships between variables in the real data and then re-construct them in synthetic data generation.
- CTGAN: Conditional Tabular Generative Adversarial Network (GAN) (Xu, Skoularidou, Cuesta-Infante, & Veeramachaneni, 2019) uses a GAN architecture to generate synthetic tabular data based on certain conditions/constraints presented in the real data. GAN is a deep learning model used for generative modelling that consists of a generator network and a discriminator network trained in an adversarial manner. Thus, CTGAN can be expected to generate synthetic data that is similar to the real data.
- CopulaGAN: It is a variation of the CTGAN that combines GaussianCopula with the CTGAN. With GaussianCopula, CopulaGAN gains the ability to capture the dependence structure between variables when learning real data.
- TVAE: Variational Autoencoder (VAE) for Tabular data (Xu et al., 2019). VAE is another deep-learning model for generative modelling that consists of an encoder and a decoder. The encoder in TVAE maps tabular data to a low-dimensional latent space representation that follows a probability distribution, while the decoder generates/re-constructs synthetic data from the latent space.

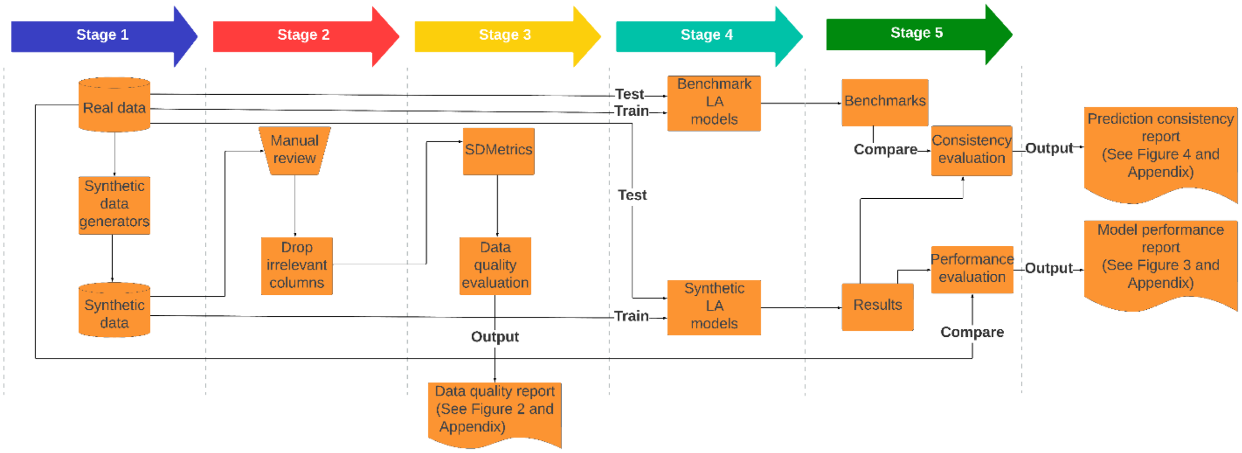
Student data

For our analysis, we used anonymised student data for three mandatory IT courses (coded as ITF, NWF, and PBS) from a large public Australian university. For each course, we used enrolments from the 2015 to 2020 academic years making up the following sample sizes: PBS (1,835), NWF (1,873), and ITF (1,829). Each course record contains students' demographic information, students' engagements with courses and other related information which we capture in the Appendix.

Experiment pipeline

Figure 1 describes the pipeline of our experiments. We used four SDGs to generate synthetic student data of the same size for three courses. To ensure stable results, we repeated the process 100 times, 50 on a device with a dedicated Graphics Processing Unit (GPU) and 50 on a device without a GPU. This allowed us to examine the impact of GPU usage on the efficiency of SDGs.

Figure 1: Flowchart of experiment pipeline



We first reviewed and pruned irrelevant columns (e.g., student IDs, columns with excessive missing values, etc.) from the synthetic datasets after generation. The Synthetic Data Metrics (SDMetrics) (MIT's Data to AI Lab, 2016) was used to assess the extent to which synthetic data resembles real data in terms of statistical properties (i.e., Column Shapes and Column Pair Trends). Column Shapes and Column Pair Trends quantify how similar the distributions of real and synthetic data are in each column and whether the synthetic data capture trends between pairs of columns that are found in the real data, respectively. We averaged the two metrics to report an overall quality score. To evaluate the utilities of synthetic student data serving LA models, we repeatedly trained LA models (linear regression and classification tree) with generated synthetic data. These models solve classic tasks in LA (i.e., course grade and outcome prediction), and are the two most commonly used methods in LA due to simplicity and interpretability (Kabra & Bichkar, 2011; Gadhavi & Patel, 2017). We evaluated the utilities of synthetic student data by comparing predictions of LA models they served against ground truth (student learning outcomes recorded in real data). Additionally, we created benchmarks of LA model performances by training and testing models with entire real data and compared the performances of LA models trained with synthetic student data with benchmarks. Consistency was further evaluated by comparing predictions of LA models trained with synthetic data versus with real data.

Results

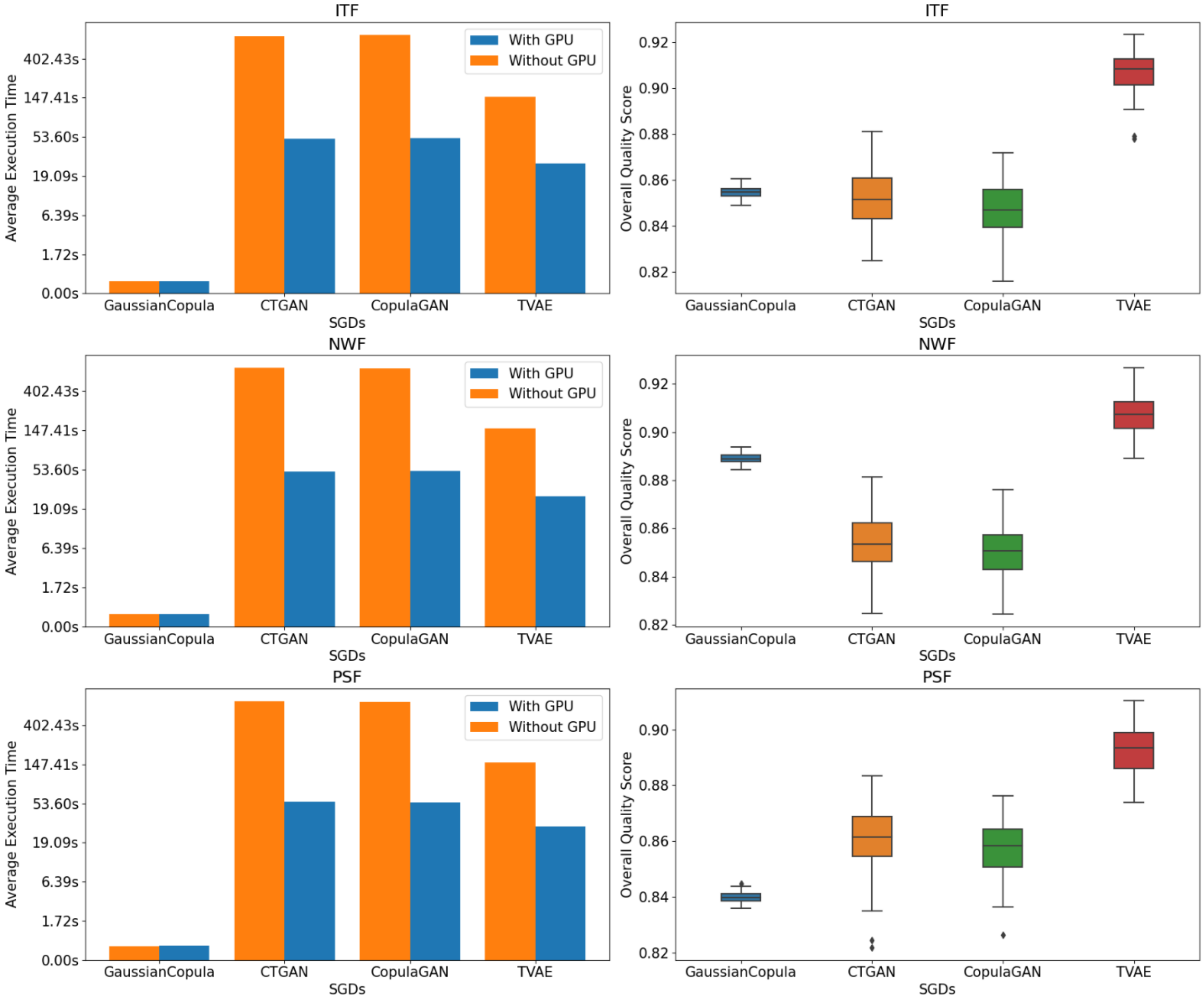
Efficiency of synthetic data generation

We evaluated the efficiencies of SDGs by comparing the execution time for synthesising the same amount of student data (with and without GPU). From Figure 2, we observed that GPU usage substantially improves the implementation efficiencies of examined SDGs (i.e., 14 times faster for CTGAN and CopulaGAN, and 6 times faster for TVAE in our case) except for the GaussianCopula. Interestingly, results showed that GaussianCopula is the fastest SDG as evidenced by less than 1 second of execution time in both with and without GPU settings. Also, we observed that data from three courses required approximate execution time for all SDGs due to similar data size.

Synthetic data quality evaluation

Figure 2 visualises overall synthetic data qualities (i.e., to what extent synthetic data is statistically similar to the real data). The GaussianCopula exhibited variable and inconsistent data quality despite its high efficiency in synthetic data generation. Results showed that its quality was comparable to GAN-based SDGs for the ITF course, superior to GAN-based SDGs for the NWF course, and was lowest for the PSF course. The data generated by TVAE have generally the highest quality across all three courses. We also observed the CTGAN and the CopulaGAN to be relatively similar in terms of the quality of the generated data. We find this not entirely surprising since the CopulaGAN is essentially a variation of the CTGAN.

Figure 2: Comparison of efficiencies of SDGs (with and without GPU) and qualities of synthetic data generated (details can be found in Appendix)

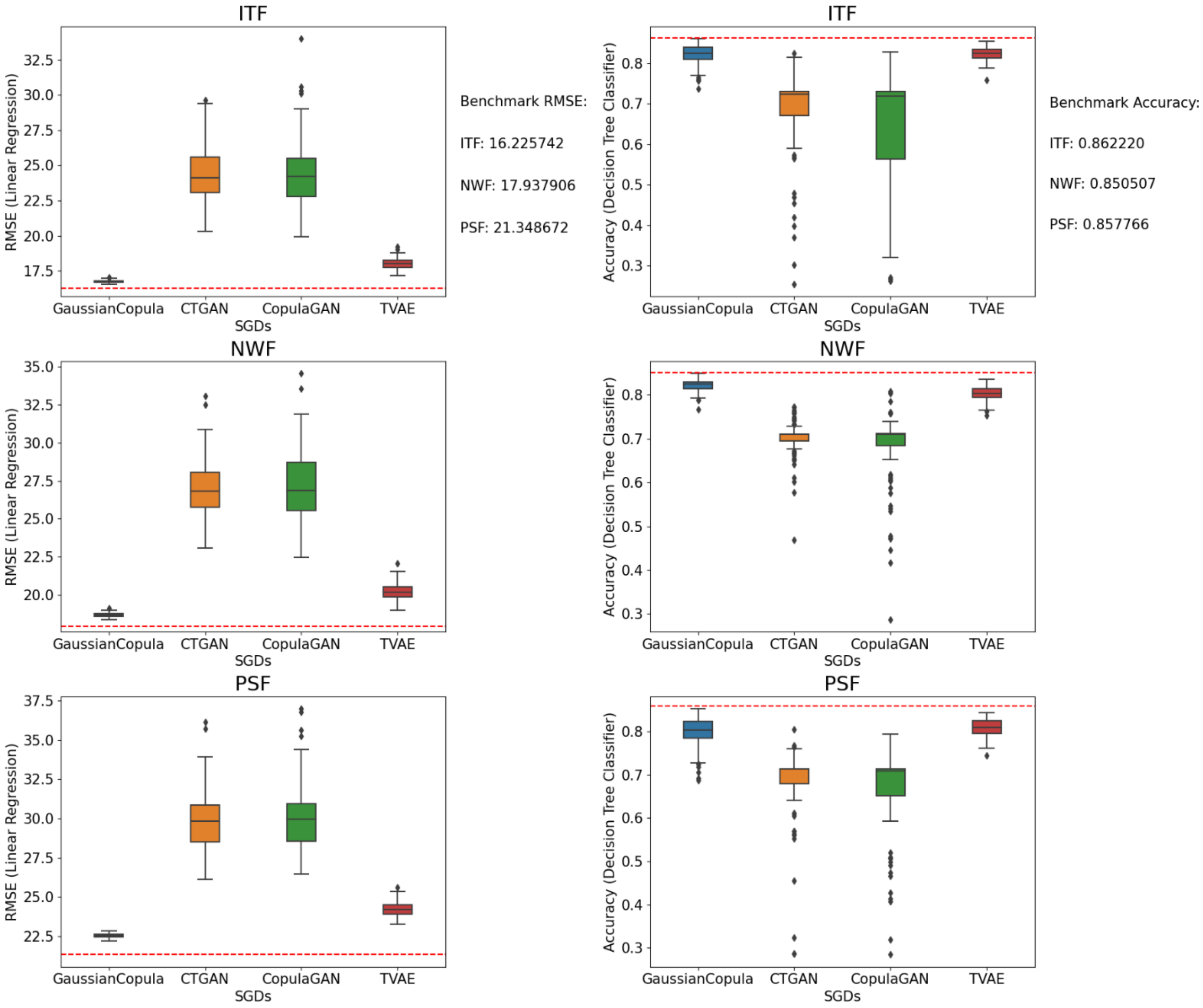


Utilities of synthetic data in serving LA models

For our referring models, we chose linear regression and the decision-tree classifier. We evaluated the performance of the linear regression using RMSE, which measures the average squared difference between the actual and predicted grades. The performance of the classifier was evaluated by accuracy (i.e., the probability of correctly identifying those who pass or fail with the passing threshold set at 50%). We presented all benchmark performances achieved from real data in the side caption in Figure 3.

We now evaluate utilities of generated synthetic data in serving LA models (hereinafter referred to using the following nomenclature: SDG name + regressor or classifier) against ground truth. Additionally, we compare the utilities of synthetic data by comparing its predictive consistency with that of real data in the learning analytics models they are used for.

Figure 3: Comparison of performances of LA models trained using synthetic data generated by SDGs (details can be found in Appendix)



Benchmark RMSE:

ITF: 16.225742

NWF: 17.937906

PSF: 21.348672

Benchmark Accuracy:

ITF: 0.862220

NWF: 0.850507

PSF: 0.857766

Boxplots depicted in Figure 3 display the distribution of specific metrics (i.e., RMSE for regressors and Accuracy for classifiers) indicating the performances of LA models. Starting with the regression models, interestingly, although the GaussianCopula appeared to generate synthetic data with unstable quality, the GaussianCopula regressors have the most similar predictive performances to the benchmark across all courses. The TVAE regressors are the second closest to the benchmark in students' grade prediction despite that TVAE-generated synthetic data achieves the highest quality score. Meanwhile, both CopulaGAN and CTGAN regressors appeared to perform inferiorly with relatively higher RMSE. With regard to classifiers, we observed that the TVAE classifiers outperformed others, as anticipated. GaussianCopula classifiers rank second across all datasets and again, the CopulaGAN and the CTGAN classifiers exhibited relatively poor performances at accurately determining students' course success.

To assess consistency between the predictions made by LA models that synthetic and real data served, we rely on Pearson Correlation (Hall, 2000) for linear regression and Cosine Similarity (Rahutomo, Kitasuka, & Aritsugi, 2012) for the classification. Results from GaussianCopula and TVAE models showed strong consistency with real LA model predictions regarding both regressor and classifier, as seen in Figure 4. These consistency trends also coincided with previous evaluations of model performance. The consistency evaluation of predictions from GAN-based regressor and classifier is fluctuating, as they have generated highly irrelevant predictions when compared to the predictions generated by the models using real data on multiple occasions.

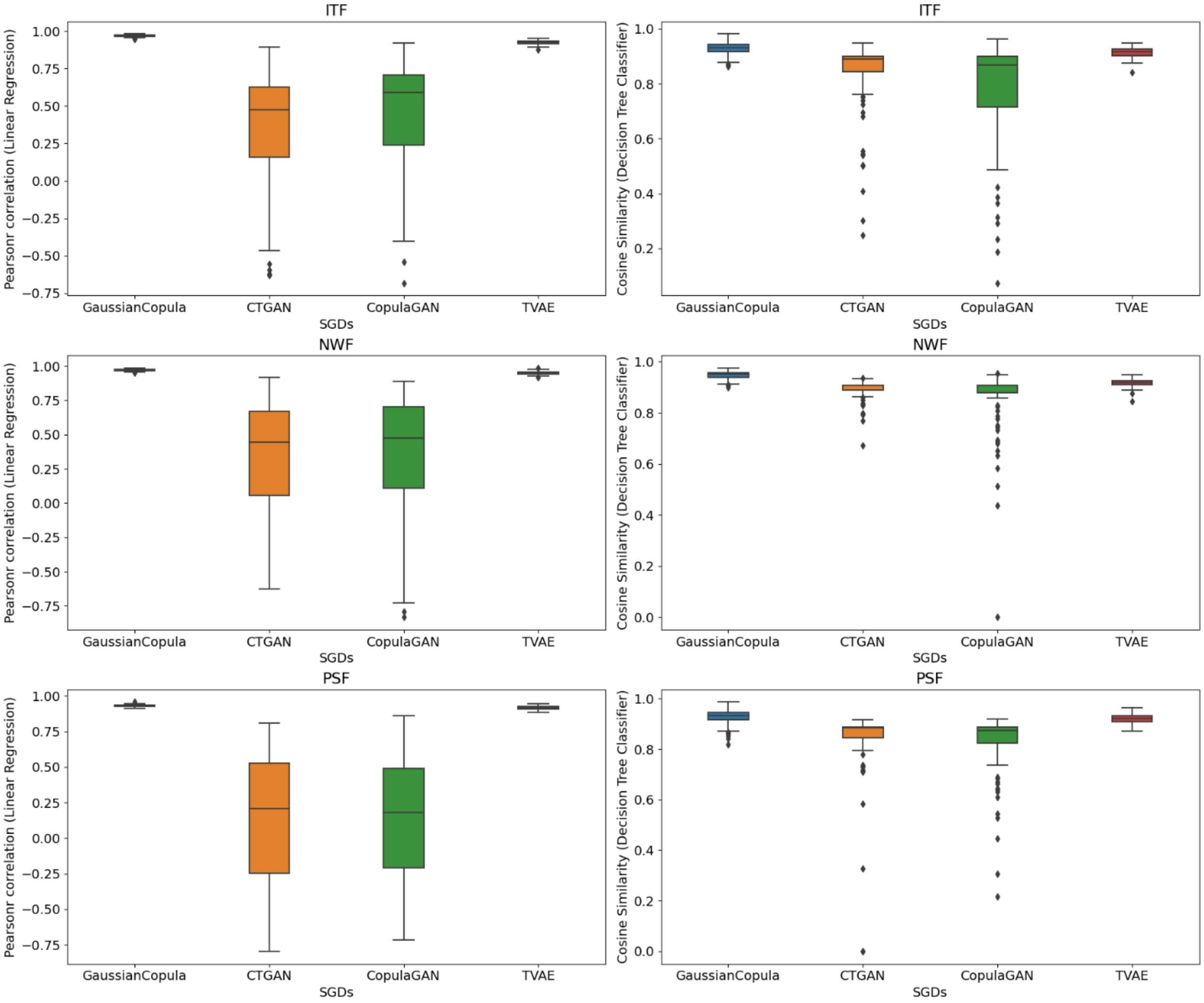
Discussion and conclusion

Deciding on the best SDG overall is not a straightforward task. It is always a trade-off choice that might depend on particular use cases and practitioners' preferences, such as the size of data, choice of LA models and whether to prioritise efficiency or quality, etc. The scope of the present comparative study is to test a range of existing SDGs, from statistical model-based to generative machine/deep learning model-based SDGs, on one of the most representative student data in order to provide guidelines and advice to researchers and practitioners in the LA community who wish to incorporate SDGs and use synthetic data in LA.

Speaking of implementation efficiency, perhaps unsurprisingly, GPU can parallelise the computation and accelerate model training for GAN-based SDGs and TVAE due to their nature of being deep learning models, while the efficiency of the GaussianCopula solely relies on CPU performance, not affected by GPU usage.

We have detailed all metrics for evaluating the statistical similarity of synthetic data against real data in the Appendix. The metrics measured how well SDGs restored value distribution (Column Shapes) and correlation between columns (Column Pair Trends) from real data to synthetic data. Our findings revealed that GaussianCopula was less effective in restoring distribution but better at capturing correlations, as expected from its model essentials. TVAE outperformed other SDGs in capturing both distribution and correlation. Acknowledged by the existing literature (Brock, Donahue, & Simonyan, 2018; Salimans et al., 2016), GAN-based SDGs introduced more randomness and diversity in data generation, leading to less stability, as indicated by the higher standard deviation in statistics and outliers in plots.

Figure 4: Prediction consistencies of LA models trained using synthetic data and real data (details can be found in the Appendix)



In terms of providing utilities to LA models, both GaussianCopula and TVAE have ensured promising performances of the LA models they served, comparable to the benchmark. Additionally, they have made consistent predictions against those made by LA models trained by real data. Particularly, GaussianCopula has presented exceptional compatibility with linear regression, possibly due to its successful preservation of correlations from real data. Notably, GaussianCopula, despite achieving the lowest average quality score for the PSF course data, continued to serve LA models adequately. This suggests that statistical similarity does not necessarily equate to good utility. GAN-based SDGs displayed varying results across performance and consistency evaluation, although certain satisfactory results were observed.

In summary, for beginners wishing to integrate SDGs into LA, we recommend GaussianCopula due to its efficiency and satisfactory performance in most use cases. TVAE is a more robust choice in all aspects, but its efficiency could be a concern without GPU. GAN-based SDGs produce diverse samples and thus are suitable for synthesising large-scale data, but require careful hyperparameter tuning and testing to overcome instability.

In this study, we focused exclusively on SDGs suitable for single table data. SDV also offers SDGs that are suitable for relational and time-series data generators, which hold promise for future applications in LA. Particularly, synthetic relational raw student data affords greater flexibility for various feature engineering approaches, which have been shown to be critical in the construction of LA models (Romero & Ventura, 2020). Nonetheless, further investigations are necessary to explore the interplay between SDGs and feature engineering approaches, in addition to LA models. Moreover, while time-series data (Esling & Agon, 2012) in LA may not be as rigorous as that in other fields such as environmental science and economics, a considerable amount of student data is timestamped, rendering its potential for temporal analysis (Shirvani Boroujeni & Dillenbourg, 2019).

With privacy and ethics in mind, we carried out this study to investigate (1) whether SDGs can replicate statistical properties of real data in generated synthetic data; and (2) whether SDGs can preserve the utility of real student data in generated synthetic data in terms of serving analytical models for common LA tasks. In conclusion, we posit that SDGs can be employed to generate synthetic data as a substitution or to complement real student data in serving LA.

Funding

No funding was received for the conduct of this research.

Disclosure statement

The authors report no potential conflict of interest.

About the authors

Chen Zhan is a data scientist and early career researcher at Centre for Change and Complexity in Learning (C3L) at the University of South Australia. He has broader experiences in conducting interdisciplinary research and collaborating with industries and governments. Recently, he has been focusing on artificial intelligence and learning analytics for the education sector. In particular, he contributes to research in the development and guidelines of responsible use AI and ethical use data in education.

ORCID: https://orcid.org/0000-0001-5723-2564

Oscar Blessed Deho received his Bachelor’s degree in Computer Science and Engineering from University of Mines and Technology, Tarkwa, Ghana in 2018. He is currently a PhD candidate in Computer and Information Science at the University of South Australia, Australia. His main research interests are fairness in machine learning, learning analytics and explainable AI.

ORCID: https://orcid.org/0000-0002-4794-8339

Xuwei Zhang received his Bachelor's degree in Internet of Things from Chongqing College of Arts and Sciences, Chongqing, China in 2017, and his Master's degree in Data Science from the University of Adelaide, South Australia, in 2023. He is currently a Research Assistant at the Centre for Change and Complexity in Learning at the University of South Australia. His main research interests are in artificial intelligence, learning analytics, and data automation processing and visualisation.

Srećko Joksimović is a Senior Lecturer in Data Science at the Education Futures, University of South Australia. His research is centred around augmenting abilities of individuals to solve complex problems in collaborative settings. Srećko is particularly interested in evaluating the influence of contextual, social, cognitive, and affective factors on groups and individuals as they solve complex real-world problems. In so doing, he utilises a wide range of methods from machine learning, artificial intelligence, and natural language processing, as well as data science and social computing in general. Srećko has been actively involved in the development of the learning analytics research field. He served on the Executive Committee of the Society for Learning Analytics Research for four years, between 2015 and 2019. Currently, Srećko is a committee member of the ASCILITE Learning Analytics SIG and a LAK21 Conference Program Chair.

ORCID: https://orcid.org/0000-0001-6999-3547

Maarten de Laat is professor in Augmented and Networked Learning and co-director of the Centre for Change and Complexity in Learning (C3L) at the University of South Australia. His research focuses on learning and value creation in social networks. He uses practice-based research methodologies to study the impact technology, AI, learning analytics and social design has on the way social networks and communities work, learn and innovate. Maarten is co-chair of the international Networked Learning Conference and editor of the Springer book series on Research in Networked Learning.

ORCID: https://orcid.org/0000-0003-2243-2667

References

Barrett, M., Branson, L., Carter, S., DeLeon, F., Ellis, J., Gundlach, C., & Lee, D. (2019). Using artificial intelligence to enhance educational opportunities and student services in higher education. *Inquiry: The Journal of the Virginia Community Colleges*, *22*(1), 11.

Benvenuti, M., & Mazzoni, E. (2020). Enhancing wayfinding in pre-school children through robot and socio-cognitive conflict. *British Journal of Educational Technology*, *51*(2), 436–458.

Berg, A. M., Mol, S. T., Kismihok, G., & Sclater, N. (2016a). The role of a reference synthetic data generator within the field of´ learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128.

Berg, A. M., Mol, S., Kismihok, G., & Sclater, N. (2016b). Scaling learning analytics: The practical application of synthetic data. In *Eden Conference Proceedings* (pp. 264–269).

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv*, 1809.11096.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Celik, I., Gedrimiene, E., Silvola, A., & Muukkonen, H. (2022, 03). Response of learning analytics to the online education challenges during pandemic: Opportunities and key examples in higher education. *Policy Futures in Education*. https://doi.org/10.1177/14782103221078401

Dorodchi, M., Al-Hossami, E., Benedict, A., & Demeter, E. (2019, December). Using synthetic data generators to promote open science in higher education learning analytics. In *2019 IEEE International Conference on Big Data* (pp. 4672-4675). IEEE.

Drachsler, H., & Greller, W. (2016). Privacy and analytics: It's a delicate issue a checklist for trusted learning analytics. In D. Gašević, G. Lynch, S. Dawson, H. Drachsler, & C. P. Rosé (Eds.*), LAK '16: Proceedings of the 6th International Conference on Learning Analytics & Knowledge* (pp. 89–98) Association for Computing Machinery.

El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: Balancing privacy and the broad availability of data*. O'Reilly Media.

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*, *45*(1), 1–34.

Flanagan, B., Majumdar, R., & Ogata, H. (2022). Fine grain synthetic educational data: Challenges and limitations of collaborative learning analytics. *IEEE Access*, *10*, 26230–26241.

Gadhavi, M., & Patel, C. (2017). Student final grade prediction based on linear regression. *Indian Journal of Computer Science and Engineering*, *8*(3), 274–279.

Hall, M. A. (2000). *Correlation-based feature selection of discrete and numeric class machine learning*. Dept. of Computer Science, University of Waikato.

Joksimović, S., Marshall, R., Rakotoarivelo, T., Ladjal, D., Zhan, C., & Pardo, A. (2022, 01). Privacy-driven learning analytics. In E. McKay (Ed.), *Manage your own learning analytics: Implement a Rasch modelling approach* (pp. 1-22). Springer. https://doi.org/10.1007/978-3-030-86316-6 1

Kabra, R., & Bichkar, R. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications*, *36*(11), 8–12.

Kavitha, V., & Lohani, R. (2019). A critical study on the use of artificial intelligence, e-learning technology and tools to enhance the learner's experience. *Cluster Computing*, *22*, 6985–6989.

MIT's Data to AI Lab. (2016). Sdmetrics. https://docs.sdv.dev/sdmetrics/

Mutimukwe, C., Viberg, O., Oberg, L.-M., & Cerratto Pargman, T. (2022). Students' privacy concerns in learning analytics: Model development. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13234

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics* (pp. 399–410).

Prinsloo, P., & Slade, S. (2016, 04). Student vulnerability, agency and learning analytics: An exploration. *Journal of Learning Analytics*, *3*, 159-182. https://doi.org/10.18608/jla.2016.31.10

Prinsloo, P., & Slade, S. (2017). Ethics and learning analytics: Charting the (un)charted. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 49-57). SoLAR.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology* (Vol. 4, p. 1).

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. *Advances in Neural Information Processing systems*, *29*.

Shirvani Boroujeni, M., & Dillenbourg, P. (2019). Discovery and temporal analysis of MOOC study patterns. *Journal of Learning Analytics*, *6*, 16-33. https://doi.org/10.18608/jla.2019.61.2

Tsai, Y.-S., Whitelock-Wainwright, A., & Gašević, D. (2020). The privacy paradox and its implications for learning analytics. In V. Kovanović, M. Scheffel, N. Pinkwart, & K. Verbert (Eds.), *LAK '20: Proceedings of the 10th International Conference on Learning Analytics & Knowledge* (pp. 230–239). Association for Computing Machinery.

Vie, J.-J., Rigaux, T., & Minn, S. (2022). Privacy-preserving synthetic educational data generation. In *Educating for a new future: Making sense of technology-enhanced learning adoption: 17th European Conference on Technology Enhanced Learning, 2022, Proceedings* (pp. 393–406).

Wan, C., Li, Z., Guo, A., & Zhao, Y. (2019). Sync: A unified framework for generating synthetic population with Gaussian copula. *arXiv*, 1904.07998.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, *32*.

Zain, S. (2021). Digital transformation trends in education. In D. Baker, & L. Ellis (Eds.), *Future directions in digital information* (pp. 223–234). Elsevier.

Appendix

Table 1: Summary of features in student data used in the present study.

Category	Features
Demographics	Age, gender, home language, citizenship status, disability status
VLE (i.e., Moodle elements) Engagement	Last login, assignment, quiz, course, forum, folder, resource, URL, page
Categorised Engagement	Informative, instructional, social
Others	Study period, SMS alerts
Target Outcome	Course grade (Regression), Course grade binarized at >= 50% (Classification)

Table 2: Efficiencies of SDGs in terms of execution time

Course Name	SDG Name	Average Time with GPU	Time Std with GPU	Average Time without GPU	Time Std without GPU
ITF	GaussianCopula	0.379266	0.009703	0.363781	0.002283
	CTGAN	51.205874	0.488692	717.671241	1.010423
	CopulaGAN	52.349525	0.322569	740.189961	1.257696
	TVAE	26.567587	0.155131	150.216883	0.175728
NWF	GaussianCopula	0.378744	0.005302	0.374642	0.001772
	CTGAN	51.284514	0.344648	726.196453	1.019346
	CopulaGAN	51.851896	0.218032	722.975016	0.985052
	TVAE	26.818652	0.264630	154.370848	0.206071
PSF	GaussianCopula	0.442902	0.007439	0.422418	0.001841
	CTGAN	56.254287	0.368436	748.628781	1.538403
	CopulaGAN	55.671634	0.281993	738.258385	1.327826
	TVAE	29.625146	0.186811	155.960989	0.230422

Table 3: Quality metrics of generated synthetic data

Course Name	SDG Name	Overall Quality Score [1]	Overall Quality Score Std	Column Shapes[2]	Column Shapes Std	Column Pair Trends[3]	Column Pair Trends Std
ITF	GaussianCopula	0.854563	0.002374	0.826392	0.002485	0.882733	0.002770
	CTGAN	0.851888	0.013275	0.860456	0.018318	0.843320	0.009590
	CopulaGAN	0.846878	0.011905	0.859163	0.016507	0.834592	0.008588
	TVAE	0.907142	0.008731	0.908748	0.008510	0.905535	0.009407
NWF	GaussianCopula	0.888991	0.001774	0.873979	0.002141	0.904003	0.001963
	CTGAN	0.854389	0.011405	0.871136	0.015554	0.837643	0.008303
	CopulaGAN	0.850116	0.010450	0.866315	0.013786	0.833917	0.008404
	TVAE	0.906967	0.007795	0.914336	0.007538	0.899598	0.009517
PSF	GaussianCopula	0.839833	0.001845	0.798342	0.002241	0.881324	0.002015
	CTGAN	0.860546	0.012687	0.865251	0.017195	0.855840	0.009514
	CopulaGAN	0.857511	0.010216	0.863863	0.013955	0.851160	0.008113
	TVAE	0.892634	0.008591	0.892601	0.009149	0.892668	0.008489

[1] Overall Quality Score averages the metrics of Column Shapes and Column Pair Trends.

[2] The Column Shapes metric describes how similar the distributions of real and synthetic data are in each column. It yields a separate score for every column. The final Column Shapes score is the average of all columns.

[3] Column Pair Trends metric quantifies whether the synthetic data capture trends between pairs of columns that were found in the real data. The trend between two columns describes how they vary in relation to each other, e.g., the correlation. It yields a score between every pair of columns and the final Column Pair Trends score averages them all.

Table 4: Performance of Linear Regression model trained by generated synthetic data and prediction consistency with outputs of LA models trained by real data using Pearson Correlation

Course Name	SDG Name	Average RMSE	RMSE Std	Average Pearson Correlation	Pearson Correlation Std
ITF	GaussianCopula	16.752500	0.106698	0.974609	0.005953
	CTGAN	24.328230	1.726677	0.362227	0.389788
	CopulaGAN	24.347140	2.466106	0.456765	0.355503
	TVAE	18.025770	0.377618	0.936392	0.012811
NWF	GaussianCopula	18.696228	0.132485	0.967597	0.006661
	CTGAN	26.953544	1.872785	0.325436	0.419433
	CopulaGAN	27.204519	2.424000	0.334115	0.438765
	TVAE	20.237349	0.536012	0.945317	0.012051
PSF	GaussianCopula	22.533960	0.133229	0.934538	0.007956
	CTGAN	29.923113	1.851817	0.125684	0.450658
	CopulaGAN	30.063285	2.205507	0.154611	0.434189
	TVAE	24.242295	0.460819	0.917711	0.011648

Table 5: Performance of Decision Tree classifier trained by generated synthetic data and prediction consistency with outputs of LA models trained by real data using Cosine Similarity

Course Name	SDG Name	Average Accuracy	Accuracy Std	Average Cosine Similarity	Cosine Similarity Std
ITF	GaussianCopula	0.819546	0.025774	0.927522	0.024249
	CTGAN	0.685970	0.101079	0.838137	0.128448
	CopulaGAN	0.644385	0.139122	0.778096	0.185133
	TVAE	0.823242	0.016703	0.913753	0.017495
NWF	GaussianCopula	0.821698	0.013795	0.947091	0.015766
	CTGAN	0.702045	0.041142	0.892618	0.037656
	CopulaGAN	0.683310	0.081315	0.861428	0.124254
	TVAE	0.803673	0.015396	0.917131	0.015471
PSF	GaussianCopula	0.798441	0.034877	0.927274	0.029793
	CTGAN	0.682131	0.083757	0.838664	0.140985
	CopulaGAN	0.664954	0.094269	0.823905	0.119547
	TVAE	0.808016	0.020098	0.920111	0.017507