Paper ET18

Synthetic Data Will Save the Day: Fact or Fiction

Sherrine Eid, MPH, SAS Institute, Inc., Cary, NC, USA
Sundaresh Sankaran, SAS Institute, Inc., Cary, NC, USA

ABSTRACT

From their early adoption in the 1980s to their current applications in genomic simulations, drug discovery, and machine learning, synthetic datasets have revolutionized research in fields such as genomics and personalized medicine.

Despite their potential, synthetic datasets present several challenges and limitations that must be addressed. Generating realistic synthetic data requires a deep understanding of the underlying biological processes, as well as sophisticated computational methods for synthesis and validation. Moreover, ensuring the quality, diversity, and representativeness of synthetic data poses significant challenges.

Methods that generate synthetic data such as Generative Adversarial Networks (GANs) or Synthetic Minority Oversampling TEchnique (SMOTE) use computationally complex methods such as deep learning and nearest neighbor search, and take advantage of scalable compute power, generating data indistinguishable from real data, without compromising subject and patient privacy.

Researchers are looking for privacy preserving, cost effective, balanced and scalable approaches to synthetic data generation.

We discuss drivers behind increased demand for, best practices and tools in synthetic data generation today.

INTRODUCTION

From their early adoption in the 1980s to their current applications in genomic simulations, drug discovery, and machine learning, synthetic datasets have revolutionized research in fields such as genomics and personalized medicine.

Despite their potential, synthetic datasets present several challenges and limitations that must be addressed. Generating realistic synthetic data requires a process paradigm – not only a deep understanding of the underlying biological processes, but also an understanding of the data lifecycle as it makes its way from source till final target data. Along the way, sophisticated computational methods for synthesis and validation are applied in order to generate a proxy which retains statistical patterns and trends, but does not expose sensitive information. Ensuring the fidelity, quality, diversity, and representativeness of synthetic data poses significant challenges.

Synthetic data is produced by purpose-built mathematical models (including generative models like GANs and VAEs) to mimic the statistical properties of real data. However, we stress that while synthetic data can accelerate machine learning research and lower costs, it is not a straightforward substitute for real data. There is an intrinsic trade-off between utility, fidelity, and privacy that must be managed carefully.

Synthetic data has tremendous potential to revolutionize early-stage research, model development, and hypothesis testing in the pharmaceutical industry by enabling secure, privacy-preserving experimentation and reducing development costs.

WHY NOW?

There are several drivers in the industry that make this discussion especially meaningful to have now. First, there is a thirst for **access to more data around the whole patient.** This increase in data demand is also associated with the growth in applications of machine learning and model training. Second, is data monetization. With the growing awareness of **opportunities to monetize and exchange data with third parties**, this is especially beneficial for Real World Data vendors and sponsors who license these data regularly for applications of Real World Evidence across the entire value chain. Especially relevant to the various privacy regulations globally, we find that there are elevated privacy risks coupled with **limitations in traditional anonymization techniques that need to be addressed**, further driving the use of more sophisticated techniques such as differential privacy in synthetic data generation. Fortunately, we now also have access to Generative AI and Agentic AI models that boosts synthetic data creation that were only theories previously.

METHODS

There exist a variety of approaches and methods to generate synthetic data. Two popular approaches are Generative Adversarial Networks (GANs) and Synthetic Minority Oversampling TEchnique (SMOTE), from which variants and implementations have also emerged.

A simplified understanding of synthetic data generation helps provide a better appreciation of GANs and SMOTE. Both these methods require real (original) data as a basis for generation. Earlier, naïve methods of generation were focused on assumptions that a variable follows an underlying distribution, and tied the generation process to that distribution (normal, Poisson etc.). Therefore, these early methods were more on the lines of data 'shaping' and did not mirror real world data closely. More *generative* methods, such as GANs and SMOTE, try to focus on preserving relationships among variables, which is important to maintain the correlational integrity of potentially hundreds of columns in a table. They approach this through different routes.

Let's look at GANs first. The simplest way to think of GANs is to play a game of "Cops and Robbers" (or Police and Robbers or any other name that you used to call it when you were a child). A GAN tries to simulate a process which essentially have two machine learning methods (formulated as neural networks) working on the same data and feeding off each other.

These two neural networks have different objective functions:

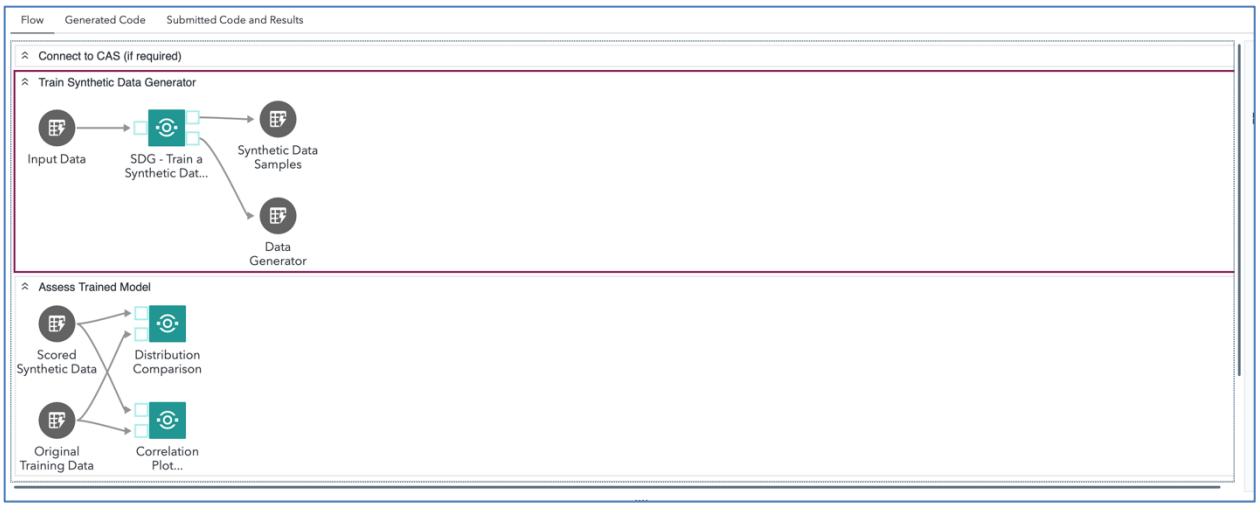
- The ***generator*** network, which executes machine learning functions yielding 'fake data'. Using real data as a base, the generator network follows an iterative process where it assigns weights at random and constructs multiple encoded representations of groups of real data (using a clustering technique), and then symmetrically distills them back to predictions of the real observations, yielding data which 'looks like' the real data, but is not exactly the same. As you may appreciate, this process is iterative in nature, and requires larger amounts of data in order to provide multiple representations as per the desired amount of observations required. This makes it a good candidate for the application of deep learning frameworks (which consist of many layers of neural networks working in interconnected fashion).

But, the generator alone is not enough to guarantee good quality synthetic data. That's where you need the 'cop' to step in. Thus, we have

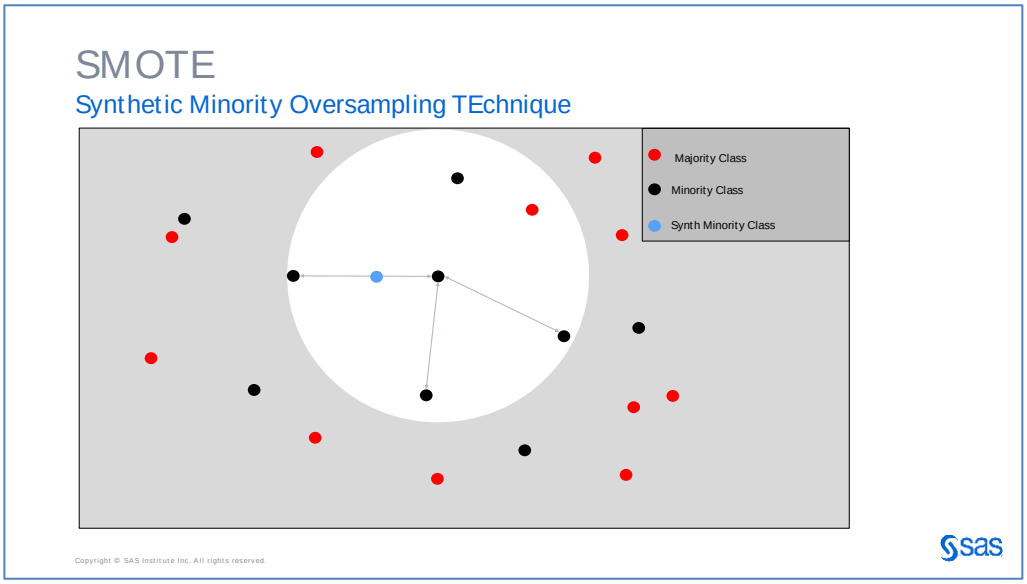
- The ***discriminator*** network, whose objective is to assess the output of the generator network and try and predict the if a record is actually fake. This is another network which operates in tandem with the generator, within the same iteration and provides back a signal regarding its confidence in the veracity (or lack thereof) of the generator's output. The generator uses this as feedback to improve its output.

The generator and discriminator networks are interdependent. Without a discriminator, the generator model might pass 'real data' as output thinking that it has done its job and terminate early. Similarly, the discriminator needs to be successfully 'fooled' into accepting a fake data observation as not fake in order to ensure quality generations. This explains the two deep learning networks operating alongside each other. The structure of the GAN also matters:- in order to ensure that variables within observations are not just generated but also maintain their relationship with each other, a Correlation Preserving Conditional Tabular GAN (CPCTGAN) is used.

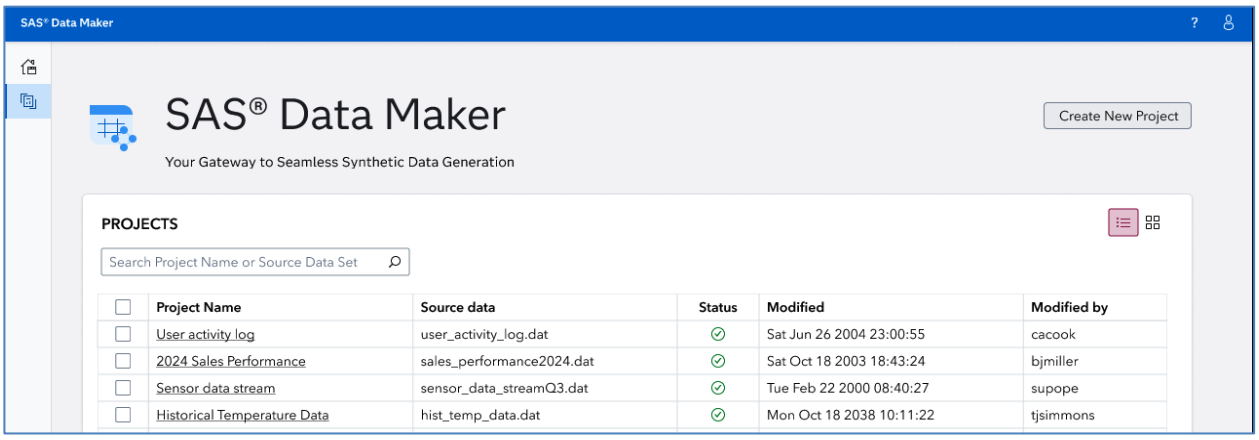
An example of a GAN implementation is the Tabular GAN procedure (PROC TABULARGAN) offered by SAS Viya, further wrapped in a convenient, low-code component called a SAS Studio Custom Step, available on GitHub (details in references). As shown in the diagram below, this also implies a separation of the training process, which yields a binary executable called a *generator model* as output, and an inferencing or scoring process, which uses the generator to create a table of observations to the requisite volume specified by the user. The benefit is that the inferencing process does not need to see the real data, and can be used by practitioners with lesser privilege to real data in the system.



On to SMOTE, next. Compared to GAN, SMOTE is a relatively easier technique which operates on the principle of nearest neighbors and does not use time-consuming deep learning frameworks. SMOTE was originally developed as a sampling technique which looks at observations which were underrepresented and tries to boost share of such observations by identifying synthetic data points located in the neighborhood of those original points. This makes it simpler to implement and requires lesser parameters, the parameters mostly relating to the number of nearest neighbors to use when identifying candidate synthetic data points. The data observations themselves may benefit from a process of transformation to be represented as higher level dimensions, which enable distance calculations on the same. Just like GANs, the SMOTE implementation is also offered on SAS Viya as an action (provided in References).



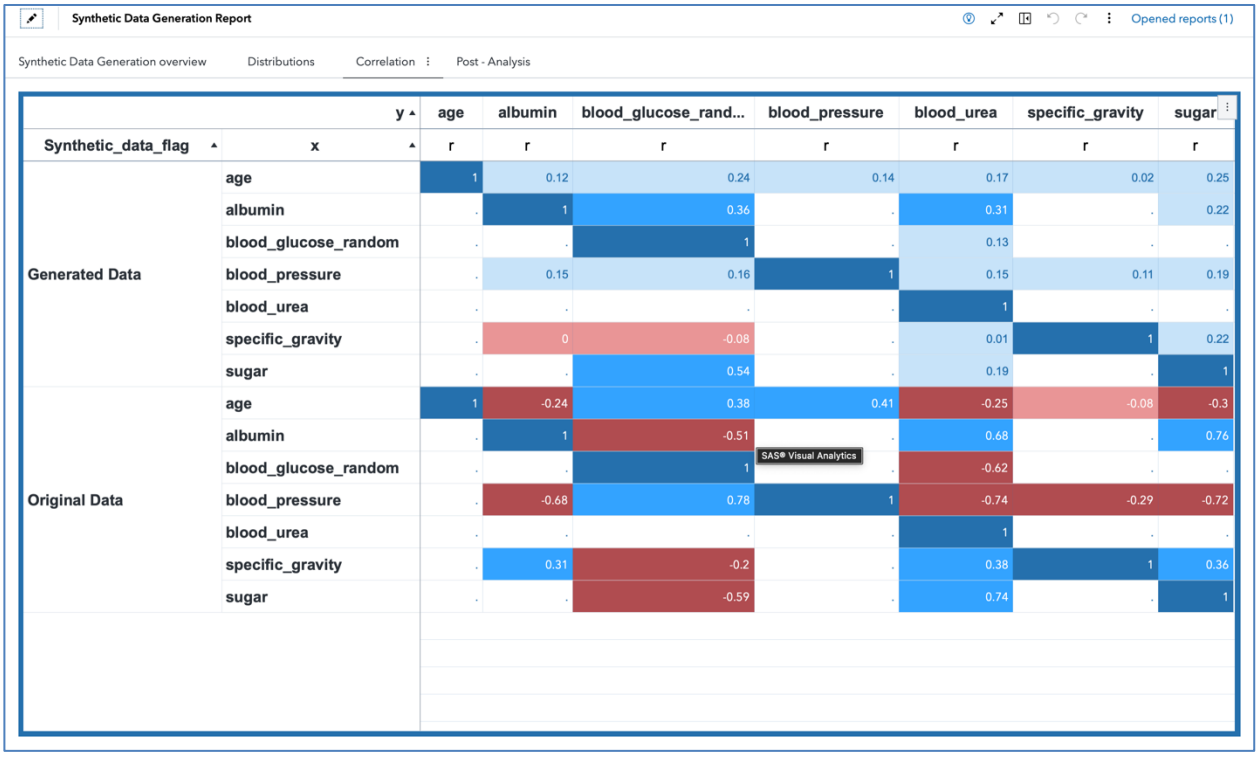
Due to strong dependencies on upstream input data quality, synthetic data generation requires a strong, project oriented structure which allows practitioners to specify configuration that can be modified to try out a number of methods, customized with specifications appropriate to input data, and also iterated and retrained where necessary. An example of this is SAS Data Maker, a offering currently under development by SAS focused on Synthetic Data Generation, which helps provide project based structures, a choice of algorithms, and the necessary layer of convenience and assessment of synthetic data.



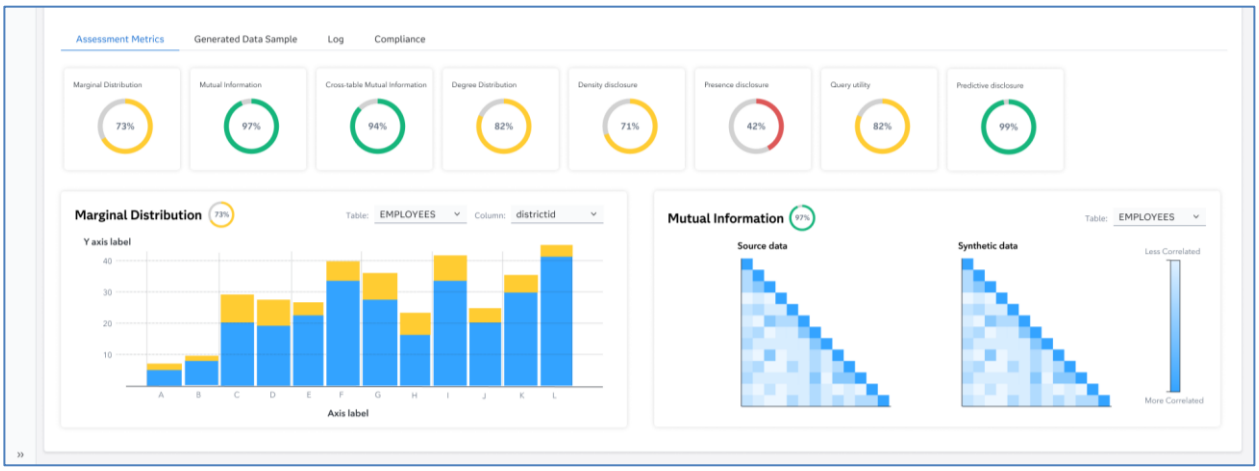
ASSESSMENT

Assessment of synthetic data helps us understand if it's 'fit for purpose', the purpose usually being a downstream predictive modeling activity, scenario analysis or similar application as decided by the practitioner. However, assessment is not a simple yes/no test because assessment is carried out on multiple dimensions which may involve tradeoffs with each other, the most important being privacy and fidelity (similarity to the original). While it may seem that a high similarity score is always desirable, drawbacks and risks exists with highly similar data, as privacy risks (specifically, re-identification risk) increase. Correspondingly, high degree of privacy might make the data 'unrealistic' and reduce its chance of providing outcomes that benefit real-life patients. Another caveat is whether any bias existent in the data gets further amplified as a result of synthetic data, which needs to be monitored and assessed.

Visualization and reports based on metrics around these aspects help decision makers assess synthetic data. Key elements of such assessment metrics include distribution comparisons between synthetic and original data for individual columns. Correlation diagrams help measure similarities and dissimilarities in inter-variable relationships of original and synthetic data. Distance measures (and metrics based on distance measures) also play a role in providing information on how 'close' synthetic data elements are to original. Practitioners may also apply simulations and attacks in an attempt to measure the extent to which privacy is protected by synthetic data.



Synthetic_data_flag	x	y: age	albumin	blood_glucose_rand...	blood_pressure	blood_urea	specific_gravity	sugar
		r	r	r	r	r	r	r
Generated Data	age	1	0.12	0.24	0.14	0.17	0.02	0.25
	albumin	.	1	0.36	.	0.31	.	0.22
	blood_glucose_random	.	.	1	.	0.13	.	.
	blood_pressure	.	0.15	0.16	1	0.15	0.11	0.19
	blood_urea	.	.	.	.	1	.	.
	specific_gravity	.	0	-0.08	.	0.01	1	0.22
	sugar	.	.	0.54	.	0.19	.	1
Original Data	age	1	-0.24	0.38	0.41	-0.25	-0.08	-0.3
	albumin	.	1	-0.51	.	0.68	.	0.76
	blood_glucose_random	.	.	1	.	-0.62	.	.
	blood_pressure	.	-0.68	0.78	1	-0.74	-0.29	-0.72
	blood_urea	.	.	.	.	1	.	.
	specific_gravity	.	0.31	-0.2	.	0.38	1	0.36
	sugar	.	.	-0.59	.	0.74	.	1



BENEFITS

We know that there are many potential benefits to utilizing synthetic data. For example, it enables **innovation** by potentially accelerating development processes by providing readily accessible data that ~~bypasses~~ does not run afoul of some regulatory and privacy hurdles. It enables rapid prototyping, model validation, and can help "unlock" data that is otherwise **hard to share**. Another benefit that excites us is the promise for privacy and fairness applications. It allows sensitive patient data to be ~~simulated~~ synthesized without directly exposing real patient records and it can help provide us an opportunity to treat historical bias in datasets. Another benefit is that it can **augment existing data** by enabling the expansion of limited datasets to improve model robustness.

Specifically to Life Sciences, we run into challenges in **imbalanced** data, evaluating and analyzing rare events, limited use of data due to regional privacy restrictions and regulations, and insufficient data for model training. Synthetic data allows us to balance datasets in a way to address and mitigate model biases that are an artifact of the data and not necessarily the model itself. Testing the robustness of rare events, adverse or otherwise, is critical for holistic model operations and hygiene. Synthetic data provides support for testing the robustness of a model's performance and handling of rare events and anomalies.

Finally, clinical trial data does not capture **sufficient** data for adequate training and testing of machine learning, deep learning and artificial intelligence models today. Generating synthetic data to execute these models is an important benefit to incorporating and leveraging "innovative" methods beyond conventional statistical methods.

Synthetic data can reduce **costs** associated with conducting clinical trials by minimizing the need for large-scale recruitment, data collection, and management. Moreover, by generating virtual patient populations and control groups, researchers can expedite trial design, simulation, and hypothesis testing, potentially accelerating the pace of drug development. Synthetic data allows researchers to explore sensitive or rare conditions without exposing real patients to potential risks, addressing ethical concerns around patient safety and consent. Furthermore, synthetic data can model diverse patient profiles and responses to treatments, facilitating personalized medicine approaches that tailor therapies to individual characteristics. Finally, researchers can use synthetic data to explore unconventional hypotheses or scenarios that may not be feasible or ethical to test using traditional methods.

RISKS

Synthetic data generation is a tool just like other analytical techniques and carries risks which have to be weighed with benefits, and which are dependent on how this tool is applied. There are several risks to utilizing synthetic data with the pharmaceutical industry despite the benefits listed above. First, there are **privacy risks** associated with generating synthetic data. Most notably, since these data are normally generated from seed, real data, information about the original data if not generated with rigorous privacy guarantees (e.g., differential privacy) could potentially be leaked. Furthermore, **outliers and low-probability events** are particularly challenging to capture privately, risking exposure of sensitive individual characteristics. A simple example of this is an individual patient who has a rare condition that might be the only patient in that region who also happens to be a member of an underrepresented demographic.

Second, there is a **tradeoff** between preserving enough statistical similarity (fidelity) to be useful for downstream tasks and ensuring that the synthetic data is sufficiently altered to protect privacy. Overemphasis on fidelity can lead to privacy breaches, while too much noise or distortion can render the data less useful. How good is good enough? Is there a compromise between fidelity and utility?

Third, we acknowledge that there may be modeling and other **technical challenges**. Complex, high-dimensional datasets, such as those in clinical trials, pose significant challenges in accurately reproducing important correlations and rare events. Also, generating synthetic data with the same representative and weighted relationships within and among tables can be a challenge for many established methods. Transparency of the generation of the data is also critical for regulatory facing work. In many instances, the "black box" nature of some generative models (e.g., GANs) can make it difficult to fully understand or trust the generated data and, therefore, may challenge the regulator's ability to approve a product with the given synthetic data generated and submitted.

This leads us to the final risk: Regulatory and ethical concerns. Today, there are no "universally accepted standards for synthetic data generation and evaluation"; it is an emerging trend. This can obviously lead to inconsistencies in how privacy, utility, and fairness are maintained. Ethical considerations remain, particularly when synthetic data is used to make critical decisions in sensitive areas like direct patient care and regulatory approval of drug products and devices that rely heavily on synthetic data.

CONCLUSION

Synthetic data offers significant promise for the pharmaceutical sector by potentially:

- Accelerating drug development and clinical trial simulations.
- Enabling research in areas with limited real-world data.
- Enhancing privacy protections for sensitive patient information.

However, it is not a magic bullet. Synthetic data must be used with caution:

- It should serve as an enabler in the early phases of research and development.
- Final models and decisions—especially those impacting patient care—should be validated and possibly fine-tuned on real data.
- Robust validation, strict privacy guarantees, and a careful balance between utility and fidelity are crucial to avoid pitfalls.

Synthetic data is a powerful tool that can help address several challenges in pharmaceutical research, but it will not completely replace the need for real data. Its success depends on responsible deployment, continuous empirical validation, and close integration with traditional data to ensure accuracy and safety in clinical applications.

REFERENCES

1. Proc tabulargan : https://go.documentation.sas.com/doc/en/pgmsascdc/default/casml/casml_tabulargan_overview01.htm

2. Smote Action: https://go.documentation.sas.com/doc/en/pgmsascdc/default/casactml/cas-smote-smotesample.htm

3. SAS Studio Custom Steps for synthetic data generation

    a. Train a GAN: https://github.com/sassoftware/sas-studio-custom-steps/tree/main/SDG%20-%20Train%20a%20Synthetic%20Data%20Generator%20through%20GANs

    b. Generate data through GANs: https://github.com/sassoftware/sas-studio-custom-steps/tree/main/SDG%20-%20Generate%20Synthetic%20Data%20through%20GANs

    c. Generate data through SMOTE: https://github.com/sassoftware/sas-studio-custom-steps/tree/main/SDG%20-%20Generate%20Synthetic%20Data%20through%20SMOTE

.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Sherrine Eid, MPH
Company: SAS Institute, Inc

Address: 100 SAS Campus Drive, Cary, NC
Work Phone: 484.294.0400
Email: Sherrine.Eid@sas.com
Website: www.linkedin.com/in/sherrineeid

Author Name: Sundaresh Sankaran
Company: SAS Institute, Inc
Address: 100 SAS Campus Drive, Cary, NC
Work Phone: 919 531 4921
Email: Sundaresh.Sankaran@sas.com
Website: www.linkedin.com/in/sundareshsankaran