Generative AI for Synthetic Data Privacy: A Framework for Tabular Data Generation and Evaluation

Under the Guidance of
Prof.Vaishali S R
Department of Data Science
Aditya College of Engineering and Technology
Bengaluru, India

Likitha B N
Department of Data Science
Aditya College of Engineering and Technology
Bengaluru, India
E-mail: likithabn2005@gmail.com

Nisarga B N
Department of Data Science
Aditya College of Engineering and Technology
Bengaluru, India
E-mail: nisargabn2@gmail.com

Thrishika P
Department of Data Science
Aditya College of Engineering and Technology
Bengaluru, India
E-mail: thrishika1306@gmail.com

Urmila C
Department of Data Science
Aditya College of Engineering and Technology
Bengaluru, India
E-mail: u961900@gmail.com

ABSTRACT

The use of data in research, machine learning, and other applications has increased the need for datasets that can be shared and analysed safely. However, real-world datasets often contain sensitive information, which makes sharing them a privacy concern. Synthetic data provides an alternative by creating artificial records based on patterns found in the original data.

This paper presents a web-based application for generating and evaluating synthetic tabular data. Users can upload CSV files, examine and preprocess their datasets, and generate synthetic records using three techniques: Conditional Tabular Generative Adversarial Networks (CTGAN), Variational Autoencoders (TVAE), and Gaussian Copula. The generated data is evaluated using SDMetrics, which examines column distributions and relationships between columns to assess statistical quality. The application also uses the NewRowSynthesis metric to identify exact matches between original and synthetic records as part of its privacy-related assessment.

The application provides visual comparisons of the original and generated datasets and allows users to download synthetic data in CSV format and evaluation reports in PDF format. The main objective is to make synthetic data generation and evaluation easier while helping users understand the balance between data quality and potential privacy risks.

Keywords— Synthetic Data, Data Privacy, Generative AI, CTGAN, Tabular Data, SDMetrics.

.

I. INTRODUCTION

The use of data-driven applications, artificial intelligence, and machine learning has increased the demand for large and diverse datasets. These datasets play an important role in training, testing, and improving machine learning models. However, real-world datasets often contain personal or sensitive information. Sharing such data for research and development can therefore create privacy concerns.

Synthetic data offers an alternative by creating artificial records based on patterns and relationships found in the original dataset. Earlier research, such as the Synthetic Data Vault (SDV), explored how statistical relationships could be used to generate synthetic records for data science applications [1]. This approach can reduce the need to directly share original records, although the generated data must still be examined for possible information leakage [7], [8].

Generating synthetic tabular data is not always straightforward. A typical dataset may contain numerical values, categorical attributes, and relationships between different columns. Xu et al. [2] addressed several of these challenges through Conditional Tabular Generative Adversarial Networks (CTGAN) and also explored TVAE for tabular data generation. Other research has examined different approaches to synthetic data generation and their applications [3], [4]. Gaussian Copula provides an additional statistical approach for modelling relationships between variables.

This project presents a web-based application that brings synthetic data generation and evaluation together in one place. Users can upload CSV datasets, examine and preprocess their data, choose a generation model, and create synthetic records. The application supports CTGAN, TVAE, and Gaussian Copula. It also uses SDMetrics to compare the original and generated datasets by examining individual column distributions and relationships between pairs of columns [5].

Along with data quality evaluation, the application includes a privacy-related assessment using the NewRowSynthesis metric

to identify exact matches between original and synthetic records [6]. Users can view comparisons between the two datasets and download the generated data and evaluation reports. The main aim of this work is to provide a practical way to explore synthetic tabular data while considering both its usefulness and potential privacy risks.

II. LITERATURE SURVEY

Several researchers have explored synthetic data generation as a way to make datasets useful for research and analysis while reducing the exposure of sensitive information. Different statistical and deep learning techniques have been developed to generate artificial records and preserve important characteristics of the original data.

Patki et al. [1] introduced the Synthetic Data Vault (SDV), a framework designed to generate synthetic data by learning statistical patterns and relationships from existing datasets. Their work showed how synthetic data could support data science applications without depending entirely on original records.

Xu et al. [2] proposed Conditional Tabular Generative Adversarial Networks (CTGAN), a technique developed specifically for generating tabular data. Their approach addresses challenges such as mixed numerical and categorical columns, uneven category distributions, and numerical values with multiple modes. The study also explored TVAE as another approach to synthetic tabular data generation.

Fonseca and Bacao [3] reviewed different methods for generating synthetic data in tabular and latent spaces. Their study discussed various generation techniques and their applications in areas such as data anonymization, data augmentation, and machine learning.

Yoon et al. [4] introduced ADS-GAN, a method that uses Generative Adversarial Networks for data synthesis in privacy-sensitive applications. Their work highlighted the importance of considering privacy when generating artificial datasets.

Apart from generation techniques, evaluating the quality of synthetic data is also important. SDMetrics provides a Quality Report that examines individual column distributions through Column Shapes and relationships between pairs of columns through Column Pair Trends [5]. These measures help compare the statistical characteristics of original and synthetic datasets.

Privacy assessment requires additional attention because generated records may still contain information related to the original dataset. The SDMetrics NewRowSynthesis metric examines whether synthetic records match records in the original data. Its documentation also recommends caution when interpreting the metric because it is a beta feature and does not provide a formal privacy guarantee [6].

Yu [7] reviewed different approaches to assessing disclosure risks in synthetic datasets, including identity, membership, and attribute disclosure. The study emphasizes that privacy evaluation should consider the possibility of information leakage rather than depending only on statistical similarity.

A consensus framework by Pilgram et al. [8] further discusses the importance of selecting suitable privacy metrics for synthetic data. It highlights that similarity-based measures alone are not sufficient to establish whether sensitive information is adequately protected.

Based on these studies, the present project brings together CTGAN, TVAE, and Gaussian Copula for synthetic tabular data generation. It uses SDMetrics to examine statistical quality and identify exact record matches as part of its privacy-related assessment. These features are provided through a single web-based application.

III. METHODOLOGY

The proposed system follows a series of steps to generate and evaluate synthetic tabular data. The process begins with uploading a dataset and preparing it for generation. Synthetic records are then created using the selected model and evaluated to understand their statistical quality and privacy-related characteristics.

A. Dataset Collection

The application allows users to upload datasets in CSV format. A sample dataset is also available for testing the application. After uploading, users can view a preview of the dataset along with basic details such as the number of rows, columns, and missing values. This helps users understand the structure of their data before proceeding.

B. Data Preprocessing

Before synthetic data generation, the uploaded dataset is examined to identify its column information and missing values. The data is prepared according to the requirements of the selected generation model. This step helps ensure that the dataset is in a suitable format for further processing.

C. Synthetic Data Generation

The application supports three techniques: Conditional Tabular Generative Adversarial Networks (CTGAN), Variational Autoencoders (TVAE), and Gaussian Copula. Users can select a model and configure the available generation settings. The selected model learns patterns and relationships from the original dataset and uses them to create artificial records.

D. Quality Evaluation

Once the synthetic records are generated, their statistical characteristics are compared with those of the original dataset using SDMetrics. The Quality Report includes Column Shapes, which examines individual column distributions, and Column Pair Trends, which examines relationships between pairs of columns. These measures help users understand how closely the generated data represents the original dataset [5].

E. Privacy-Related Assessment

The application uses the NewRowSynthesis metric to identify exact matches between records in the original and synthetic datasets [6]. It also displays a privacy-related indicator based on the New Row Rate:

Exact-Match Rate = 1 − New Row Rate

This value indicates the proportion of synthetic records that match records in the original dataset, provided the New Row Rate represents the proportion of novel records. It is only an indicative measure and does not establish complete privacy protection or provide a formal differential privacy guarantee.

F. Result Visualization and Download

After evaluation, the application displays the generated records, quality metrics, and visual comparisons between the original and synthetic datasets. Users can examine the results and download the generated dataset in CSV format. The evaluation report is also available for download in PDF format.

IV. EXISTING SYSTEM

Existing approaches to synthetic data generation use statistical techniques and machine learning models to create artificial datasets from real-world data. These methods are commonly explored in research, data analysis, and machine learning applications. However, generating useful synthetic data and checking its quality and privacy-related risks can involve several separate steps.

Traditional statistical methods generate artificial records based on predefined distributions and relationships between variables. Although these methods can work well with simpler datasets, they may not capture complex patterns, especially when the data contains different types of attributes.

Deep learning techniques, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), provide other ways to learn patterns from existing datasets. However, applying these techniques to tabular data can be challenging because numerical and categorical values may have different distributions and complex relationships.

Another important concern is the evaluation of generated data. A synthetic dataset may appear statistically similar to the original dataset, but this alone does not establish that it is suitable for every application or that sensitive information is protected. Privacy-related risks must also be examined because generated records may reveal information about the original data [7], [8].

In many existing workflows, data preprocessing, synthetic data generation, quality evaluation, and privacy assessment are handled using different tools or programming procedures. This can make the overall process difficult for users who do not have extensive technical knowledge.

These limitations create a need for an integrated approach that brings different tabular data generation techniques and evaluation methods together in a single application.

V. PROPOSED SYSTEM

The proposed system is a web-based application developed to make synthetic tabular data generation and evaluation easier. Instead of using separate tools for each stage, the application brings dataset uploading, preprocessing, synthetic data generation, and evaluation together in one platform. It allows users to generate artificial records and examine how they differ from the original dataset.

A. Key Components

The application consists of the following main modules:

1. **Dataset Collection Module:** Allows users to upload datasets in CSV format or choose a sample dataset provided by the application. It also displays a preview and basic information about the uploaded data, such as the number of rows, columns, and missing values.
2. **Data Preprocessing Module:** Examines the uploaded dataset and prepares it for synthetic data generation. It checks column information and missing values so that users can understand the structure of the data before model training.
3. **Synthetic Data Generation Module:** Provides three generation techniques: CTGAN, TVAE, and Gaussian Copula. Users can select a model and configure the available settings to generate artificial records based on patterns learned from the original dataset.
4. **Quality Evaluation Module:** Uses SDMetrics to compare the statistical characteristics of the original and synthetic datasets. It examines individual column distributions through Column Shapes and relationships between columns through Column Pair Trends.
5. **Privacy Assessment Module:** Uses the NewRowSynthesis metric to identify exact matches between original and synthetic records. The application also displays a privacy-related indicator based on the New Row Rate to help users examine potential record-level overlap.
6. **Output Module:** Displays the generated records, evaluation metrics, and visual comparisons. It also provides options to download the synthetic dataset in CSV format and the evaluation report in PDF format.

B. System Workflow

The process begins when a user uploads a CSV dataset or selects a sample dataset from the application. The uploaded data is then examined and prepared for generation. Next, the user selects one of the available models and configures the generation settings. The selected model creates synthetic records using the patterns learned from the original data.

After generation, the application evaluates the synthetic dataset using SDMetrics. The evaluation results include statistical quality measures and privacy-related information based on exact record matches. Finally, the results are displayed through the application interface, allowing users to explore the generated data, compare it with the original dataset, and download the available outputs.

C. Implementation Technologies

The application is developed using Python and Streamlit. The Synthetic Data Vault (SDV) library provides the generation techniques, including CTGAN, TVAE, and Gaussian Copula. SDMetrics is used to evaluate the quality of the generated data and examine exact record matches. Pandas supports tabular data processing, while Plotly is used to create visual comparisons. The application also supports CSV and PDF downloads.

D. Evaluation Approach

The generated dataset is examined from two perspectives: statistical quality and privacy-related characteristics. The Quality Report compares column distributions and relationships between pairs of columns. The

NewRowSynthesis metric identifies exact record matches between the original and synthetic datasets. Together, these measures provide users with a basic understanding of the generated data and its limitations. However, they should not be treated as proof of complete privacy protection.

Implementation Acknowledgement: The development of this application was initiated using the open-source Synthetic Dataset Generator project available on GitHub. The original repository served as a starting point for further development and customization. The present work focuses on bringing synthetic data generation techniques and evaluation features together within the application.

VI. ADVANTAGES

The proposed application offers several useful features that make synthetic tabular data generation and evaluation easier to understand and perform.

1. **Multiple Generation Techniques:** The application supports CTGAN, TVAE, and Gaussian Copula. Users can explore different techniques and observe how each one generates synthetic records.
2. **Easy-to-Use Interface:** The web-based interface allows users to upload datasets, select generation models, and view results without having to perform every step manually through separate programs.
3. **Statistical Quality Evaluation:** SDMetrics helps users compare the original and synthetic datasets by examining column distributions and relationships between columns.
4. **Privacy-Related Assessment:** The application includes the NewRowSynthesis metric to identify exact matches between original and generated records. This helps users examine one aspect of potential privacy risk.
5. **Visual Comparison:** Users can view graphical comparisons of the original and synthetic datasets, making it easier to notice similarities and differences in the generated data.
6. **Convenient Data Export:** The application allows users to download generated datasets in CSV format and evaluation reports in PDF format for further use and reference.
7. **Integrated Workflow:** Dataset uploading, preprocessing, generation, evaluation, visualization, and downloading are available within a single application. This reduces the need to switch between multiple tools.

These advantages describe the features available in the implemented application. However, the evaluation measures do not guarantee complete privacy protection, and the quality of generated data may vary depending on the dataset and generation technique.

VII. FUTURE ENHANCEMENTS

The current application provides a basic platform for generating and evaluating synthetic tabular data. However, there are several areas where it can be improved to make the system more reliable, efficient, and useful for different types of datasets.

1. **Improved Data Preprocessing:** Future versions can include better methods for handling missing values, outliers, different data types, and complex categorical attributes. This can help prepare datasets more effectively before generation.
2. **Advanced Generative Models:** Additional generative models and improved training techniques can be explored to produce more realistic synthetic datasets and handle a wider range of data patterns.
3. **Enhanced Privacy Evaluation:** The application can be extended with more advanced privacy assessment methods, such as membership inference and attribute disclosure testing. These methods can provide a better understanding of possible information leakage.
4. **Support for Larger Datasets:** The system can be optimized to process larger and more diverse datasets. Improvements in memory usage and processing time would make it more suitable for practical applications.
5. **Model Performance Comparison:** Further experiments can be conducted using common datasets to compare CTGAN, TVAE, and Gaussian Copula. Comparing their quality, generation time, and privacy-related evaluation results would help users understand the strengths and limitations of each technique.
6. **Improved Visualization:** More charts and comparison options can be added to help users understand differences between the original and synthetic datasets more clearly.
7. **Scalability and Performance:** The application can be improved to reduce model training time and handle larger workloads efficiently. This would make the system more convenient when working with complex datasets.
8. **Extended Report Generation:** Future versions can provide more detailed downloadable reports containing evaluation metrics, statistical comparisons, and information about the selected generation model and its configuration.

These improvements could make the application more flexible and useful while providing a broader understanding of synthetic data quality and privacy-related risks.

VIII. CONCLUSION

This paper presents a web-based application for generating and evaluating synthetic tabular data using different data generation techniques. The application supports CTGAN, TVAE, and Gaussian Copula, allowing users to upload datasets, prepare their data, and generate artificial records through a single platform.

The generated datasets are evaluated using SDMetrics to examine column distributions and relationships between different attributes. The application also uses the NewRowSynthesis metric to identify exact matches between original and synthetic records. Along with these evaluation features, users can view graphical comparisons and download the generated datasets and evaluation reports.

The main contribution of this work is bringing synthetic data

generation and evaluation together in an accessible web-based application. It provides a practical way to explore different generation techniques and understand how the generated data compares with the original dataset.

However, generating synthetic data does not automatically guarantee privacy. The evaluation methods used in this application provide only limited information about potential disclosure risks and do not offer a formal differential privacy guarantee. Further improvements and experimentation are needed to enhance data quality, generation performance, and privacy assessment.

IX. REFERENCES

[1] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410, doi: 10.1109/DSAA.2016.49.

[2] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data using Conditional GAN," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 7333–7343.

[3] J. Fonseca and F. Bacao, "Tabular and latent space synthetic data generation: a literature review," *Journal of Big Data*, vol. 10, article 115, 2023, doi: 10.1186/s40537-023-00792-7.

[4] J. Yoon, L. N. Drumright, and M. van der Schaar, "Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2378–2388, 2020, doi: 10.1109/JBHI.2020.2980262.

[5] SDMetrics, "Quality Report," *SDMetrics Documentation*. [Online]. Available: https://docs.sdv.dev/sdmetrics/data-metrics/quality. Accessed: Nov. 15, 2026.

[6] SDMetrics, "NewRowSynthesis," *SDMetrics Documentation*. [Online]. Available: https://docs.sdv.dev/sdmetrics/data-metrics/metrics-in-beta/newrowsynthesis. Accessed: Nov. 15, 2026.

[7] Z. S. Yu, "Synthetic Data Disclosure Risk Assessment: A Literature Review," in *Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics*, Statistics Canada, 2025.

[8] L. Pilgram et al., "A consensus privacy metrics framework for synthetic data," *Patterns*, vol. 6, no. 10, article 101320, 2025, doi: 10.1016/j.patter.2025.101320.