Project Summary

Batch details	PGP in Data Science with Specialization in GenAI
Team members	Shreya Bhat Bento M J Fernandes Shristi Agrawal Chinmayee Y Venkatesh Pattar
Domain of Project	Generative AI
Proposed project title	Generative AI for test Data synthesis
Group Number	6
Team Leader	Shreya Bhat
Mentor Name	Vaibhav Kulkarni

Date:

Signature of the Mentor

Signature of the Team Leader

Table of Contents

Project Details

OVERVIEW:

- This project focuses on generating synthetic test data using Generative AI techniques.
- Many industries lack large, diverse, and privacy-safe datasets for robust testing and ML training.
- The solution uses GANs, VAEs, Diffusion Models, and LLMs to create realistic synthetic datasets.
- This approach helps improve test coverage, reduce dependency on sensitive data, and accelerate development cycles.
- The end goal is to provide high-quality, scalable, and privacy-preserving test data across different domains.

Business problem statement

1. Business Problem Understanding

- Organizations depend on restricted or sensitive real data for testing.
- Limited access affects test coverage, increases errors, and causes compliance challenges.
- Testing teams often face delays due to lack of clean, diverse, and complete datasets.

2. Business Objective

- To generate scalable, realistic, and privacy-preserving synthetic test data.
- Ensure availability of high-quality datasets for testing, analytics, and ML development.
- Allow teams to test edge cases and rare-event scenarios effectively.

3. Approach

- Using Generative AI models like:
  - Tabular GANs
  - Variational Autoencoders (VAEs)
  - Diffusion Models
  - LLM-based text generators

- Evaluating the model using:
  - SDMetrics
  - KS-test
  - Correlation Preservation
  - TSTR (Train on Synthetic, Test on Real)

4. Conclusions

- Generative AI provides realistic test data without privacy risks.
- Enhances software testing accuracy and ML model reliability.
- Reduces dependency on sensitive real-world datasets.
- Improves development speed and product quality.

TOPIC SURVEY IN BRIEF

1. Problem Understanding

Industries like finance, healthcare, e-commerce, and telecom depend on high-quality data for testing systems, building ML models, and running analytics.

However, access to real data is restricted due to privacy laws (GDPR, HIPAA), internal policies, and security concerns. Additionally, real datasets may be incomplete, imbalanced, or lacking rare edge cases that are crucial for robust testing.

2. Current Solution to the Problem

Traditional solutions include manual test data creation, anonymization techniques, and rule-based data generators. While useful, anonymization often leads to data utility loss, and rule-based generators fail to capture complex real-world relationships. These methods cannot generate rare events, realistic noise patterns, or domain-specific distributions required for modern ML systems. As a result, testing is incomplete and prone to production failures.

3. Proposed Solution

Generative AI models learn the statistical distribution of available datasets and synthesize new data that resembles real data without copying original records. Models such as Tabular GANs, VAEs, Copula-based models, TimeGAN (for time series), and LLM-based synthetic text generators can produce rich test datasets with high fidelity. This synthetic data can be used for software testing, ML model training, performance evaluation, and scenario simulation.

4. **Reference to the Problem**
   - Enterprises in banking, healthcare, e-commerce, and telecom increasingly rely on synthetic data.
   - MIT, Gartner, Google, and industry research highlight synthetic data as a key technology for privacy and AI development.
   - Synthetic data addresses regulatory restrictions and testing challenges.

CRITICAL ASSESSMENT OF TOPIC SURVEY

1. **Gaps Identified**
   - Traditional anonymization loses data quality and relationships.
   - Rule-based generation lacks complexity and realism.
   - Current solutions do not provide scalable or privacy-safe datasets.
   - Limited ability to generate rare scenarios or maintain correlations.

2. **Key Gaps the Project Solves**
   - Produces high-quality synthetic data while preserving statistical structure.
   - Generates rare events and edge cases for stronger testing.
   - Eliminates privacy risks associated with real data usage.
   - Provides scalable and domain-flexible test data generation for enterprises.

METHODOLOGY TO BE FOLLOWED

1. Data Understanding & Preprocessing

- Collecting the datasets from UCI and Government Open Data portals.
- Perform cleaning (null handling, outlier removal, encoding, scaling).
- Identify data type: tabular / text.
- Split real data into training sets for generative models.

2. Model Selection

Depending on data type:

a. Tabular Data

- CTGAN / TVAE (by SDV)
- Gaussian Copulas
- TabDDPM (Tabular Diffusion Model)

b. Text Data

- LLMs (GPT, LLaMA, Mistral) for generating realistic text, emails, logs, etc.

3. Model Training

- Train generative models to learn the underlying probability distribution of the original dataset.
- Use loss functions such as:
  - Adversarial Loss (GANs)
  - KL Divergence (VAEs)
  - Reconstruction Loss
  - Diffusion Noise Prediction Loss

4. Synthetic Data Generation

- Generate synthetic samples equal to or larger than the original dataset.
- Ensure class balance and rare-event upsampling.
- Allow configurable generation volume for different test scenarios.

5. Evaluation & Validation

Use a combination of:

a. Statistical Metrics

- Kolmogorov–Smirnov (KS-Test)
- Chi-square test
- Jensen-Shannon Divergence
- Correlation Preservation Score

b. ML-based Metrics

- TSTR (Train on Synthetic, Test on Real)
- TRTS (Train on Real, Test on Synthetic)

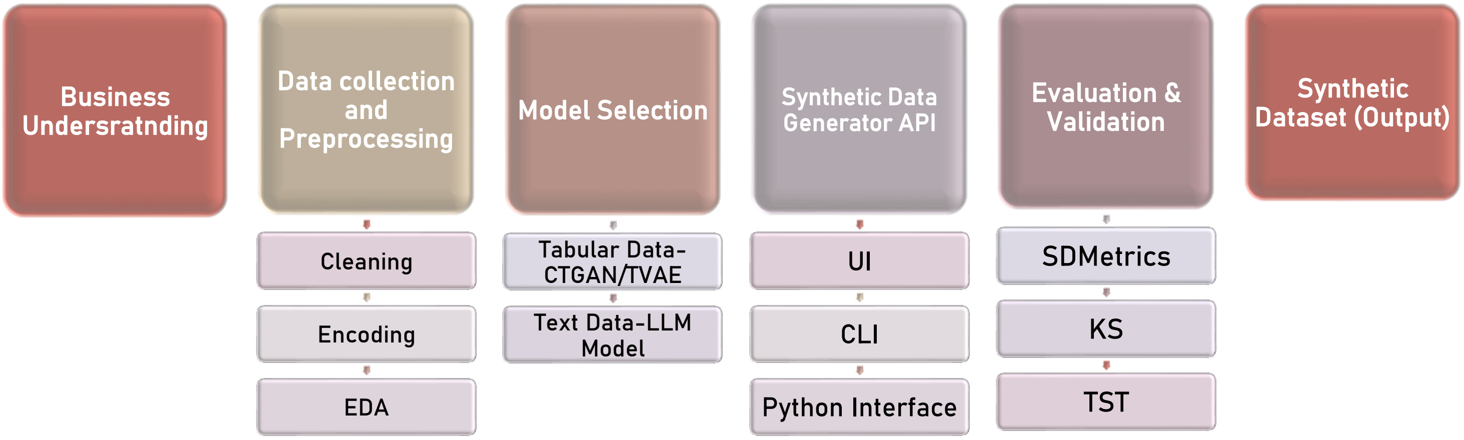
c. Privacy Metrics

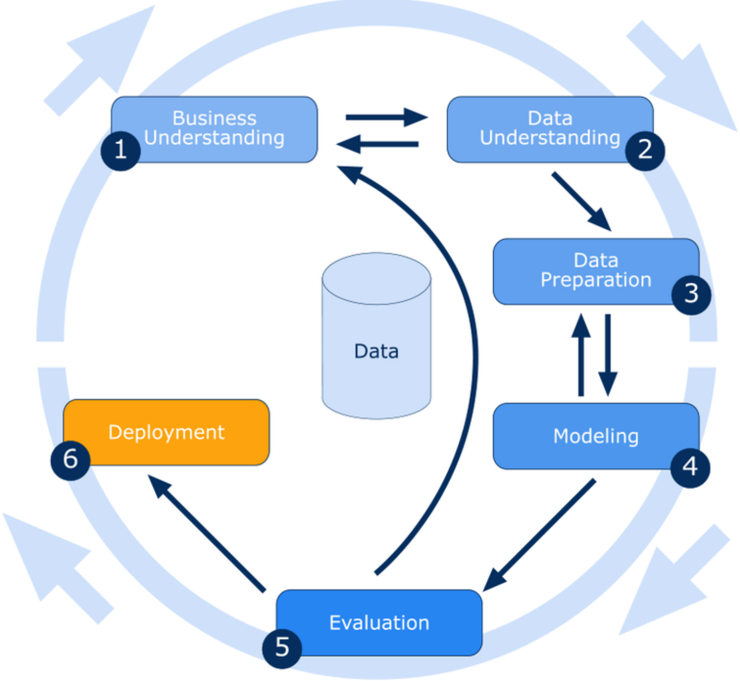
- Distance-to-closest-record
- Attribute disclosure risk
- Membership inference detection

6. Deployment of Synthetic Data Generator

- Add a UI or API layer for generating new test datasets on demand.
- Support exporting data in CSV/JSON/Parquet.
- Ensure reproducibility via model versioning and seed control

Methodology Flow





REFERENCES

1. **Salesforce (Blog) — *What Is Synthetic Data?***
   Explains how generative AI can produce realistic synthetic datasets, including rare "what-if" scenarios, safely without exposing private info. Salesforce
2. **SAP Blog — *How Synthetic Data Can Prevent AI Bias and Cyber Threats***
   Shows how synthetic data helps preserve privacy, prevents bias, and reduces legal or cyber risk when using real data. SAP
3. **EY India — *Synthetic Data: Fake Is the New Real***
   Discusses how synthetic data accelerates product testing, supports AI when data is scarce, and helps with compliance. EY
4. **TechTarget — *What Is Synthetic Data? Examples, Use Cases and Benefits***
   Detailed article covering real-world advantages: customizable data, cost-effectiveness, scaling, and use in testing & ML-training. TechTarget
5. **IndiaAI.gov.in — *Synthetic Data: Description, Benefits and Implementation***
   Indian government–oriented perspective: synthetic data helps resolve privacy issues, speeds up testing, and aids regulated domains (healthcare, banking, etc.). IndiaAI

Sample Reference for Datasets

Original owner of data	U.S. Census Bureau – extracted from the 1994 and 1995 Current Population Survey (CPS).
Data set information	The Census Income (Adult) dataset contains demographic and employment-related attributes for individuals, with the goal of predicting whether a person's income exceeds $50K annually. It includes 48,842 instances and 14 attributes such as age, workclass, education, occupation, race, hours-per-week, etc. The dataset is widely used for classification tasks, fairness analysis, model benchmarking, and synthetic data generation.
Any past relevant articles using the dataset	1. “Classification and Regression by Combining Models” – Kohavi & Becker (Used for benchmark testing). 2. “Fairness and Machine Learning” research papers commonly use this dataset to evaluate demographic bias. 3. Many Synthetic Data research papers (e.g., CTGAN, SDV) use this dataset as a standard benchmark for evaluating generative models. 4. The dataset is also used in numerous ML tutorials on income prediction and bias detection.
Reference	Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
Link to web page	https://archive.ics.uci.edu/dataset/20/census+income

Appendix

i.	AI	Artificial Intelligence
ii.	API	Application Programming Interface
iii.	CPS	Current Population Survey
iv.	CSV	Comma-Separated Values
v.	CTGAN	Conditional Tabular GAN
vi.	EDA	Exploratory Data Analysis
vii.	GAN	Generative Adversarial Network
viii.	GDPR	General Data Protection Regulation
ix.	GPT	Generative Pretrained Transformer
x.	HIPAA	Health Insurance Portability and Accountability Act
xi.	JSON	JavaScript Object Notation
xii.	KL	Kullback–Leibler Divergence
xiii.	LL.M / LLM	Large Language Model
xiv.	ML	Machine Learning
xv.	SAP	Systems, Applications & Products in Data Processing
xvi.	SDV	Synthetic Data Vault
xvii.	SDMetrics	Synthetic Data Metrics Framework
xviii.	KS-Test	Kolmogorov–Smirnov Test
xix.	TSTR	Train on Synthetic, Test on Real
xx.	TRTS	Train on Real, Test on Synthetic
xxi.	TVAE	Tabular Variational Autoencoder
xxii.	UI	User Interface

****