Capstone Project: Synthetic Data Generation

Objective

Develop a synthetic data generation module for creating high-quality, domain-relevant datasets for testing and model training.

Description

This Proof of Concept (POC) focuses on generating realistic synthetic data that preserves the statistical properties, correlations, and patterns of real-world datasets while ensuring privacy, compliance, and reproducibility. The generated data should mimic real-world data distributions and support multiple data formats (CSV, JSON, SQL, or Parquet). The project aims to enable organizations to train, validate, and stress-test AI models and analytics systems in data-constrained or privacy-sensitive environments.

Key Focus Areas

1. **Data Profiling and Schema Detection** – Analyze real datasets to detect structure, constraints, distributions, and interdependencies.
2. **Synthetic Data Modeling** – Use probabilistic models or generative AI (e.g., GANs, VAEs, LLM-based tabular generators) to simulate realistic datasets.
3. **Privacy Preservation Techniques** – Implement differential privacy or k-anonymity techniques to ensure compliance with data protection regulations.
4. **Data Validation and Quality Metrics** – Compare synthetic data to original datasets using similarity measures (KL divergence, correlation scores, etc.).
5. **Domain Adaptation** – Incorporate domain rules or business constraints to generate contextually accurate data (e.g., realistic financial transactions or healthcare records).
6. **Configurability and Automation** – Enable customizable generation (data size, features, statistical rules) and automated dataset creation.

Example Use Cases

- **Training data for AI models in regulated domains** such as finance, healthcare, or pharma.
- **Data augmentation** for low-sample ML scenarios to improve model generalization.
- **Stress-testing analytics pipelines** to evaluate system robustness under diverse data conditions.

Expected Outcome

A configurable synthetic data generation system capable of producing high-quality, representative, privacy-safe datasets aligned with specific business domains. The system should allow interactive customization and output validation, ensuring datasets are statistically consistent and compliant for AI/ML use cases.

Supportive Guide: Implementation Hints

1. **Environment Setup**
2. Install necessary libraries: `pandas`, `scikit-learn`, `sdv`, `faker`, `numpy`, and `matplotlib`.
3. Configure a workspace in Jupyter Notebook or Google Colab.
4. **Data Profiling**
5. Use `pandas_profiling` or `ydata-profiling` to explore data distributions.
6. Identify categorical, numerical, and date features.
7. **Model Selection for Data Generation**
8. Use **SDV (Synthetic Data Vault)** or **CTGAN** for complex tabular data.
9. For simpler datasets, employ `faker` for rule-based synthetic data.
10. **Privacy Controls**
11. Add noise using differential privacy methods.
12. Mask sensitive identifiers (PII) before generation.
13. **Data Validation**
14. Compare distributions of synthetic vs. real data using histograms or correlation matrices.
15. Calculate fidelity metrics (e.g., Jensen-Shannon divergence).
16. **Automation and Scalability**
17. Wrap the pipeline into a Python module for repeatable generation.
18. Integrate with cloud storage or APIs for data delivery.
19. **Visualization and Reporting**
20. Generate visual reports comparing real and synthetic data.
21. Provide configuration summaries (generation parameters, privacy scores, etc.).

Deliverables: - Synthetic Data Generation Notebook or Script - Data Quality Comparison Report - Privacy and Compliance Checklist - Configuration Template for Reproducibility