**Project Title:** AutoSynth – Autonomous Synthetic Data Generator

**Tagline:** An AI agent-based system that intelligently generates structured, high-quality synthetic tabular datasets for ML development, testing, and privacy-preserving simulation.

**🌟 Project Vision**

AutoSynth aims to solve the often-overlooked bottleneck in data-centric AI: the **lack of clean, domain-relevant, and diverse structured data**. Instead of relying on prebuilt datasets or randomly generated data, AutoSynth uses a modular architecture of AI agents to **understand, generate, validate, and evolve** synthetic tabular datasets with realistic properties.

**🔍 Problem Statement**

Most ML teams struggle with:

* Lack of high-quality data due to privacy or access restrictions
* Manual and repetitive test dataset creation
* Poor coverage of edge cases and anomalies in training data
* Inability to simulate realistic data drifts, correlations, or rare events

AutoSynth solves this by creating **domain-aware**, **logic-compliant**, and **automatically evolving** synthetic datasets.

**⚙️ System Architecture Overview**

**1. Schema Agent**

* Parses user-defined schema or prompt-generated schema
* Identifies types (numerical, categorical, ordinal, datetime)
* Input : Takes CSV, natural language prompt
* Output : Gives out the schema in json format

**2. Generator Agent**

* Uses models like CTGAN, TVAE, or Gaussian Copula to generate rows
* Supports conditional generation and rare-pattern synthesis

**3. Validator Agent**

* Applies domain constraints, logic rules, type checks
* Flags or corrects invalid records (optional rule-based or learned models)

**4. Evolution Agent**

* Introduces mutations, edge cases, and data drift scenarios
* Supports rare outlier generation, class imbalance simulation

**5. Profiler Agent**

* Evaluates dataset coverage, diversity, entropy, correlation structure
* Provides statistical summaries and warnings

**6. Logger Agent**

* Logs all generation parameters, anomalies, and evolution metadata

**🚀 Core Features**

* **Schema-aware generation** (auto-generated or user-defined)
* **Agent-driven modular pipeline**
* **Support for domain prompts** to generate schema and constraints
* **GAN/VAE-based tabular data generation**
* **Rule- and model-based validation**
* **Edge case and anomaly simulation**
* **Profile reports** to track distribution and data health

**📊 Target Use-Cases**

* ML model benchmarking
* Testing under rare scenarios
* CI pipelines for model validation
* Privacy-preserving training data generation
* Data drift and class imbalance testing

**👩‍💻 User Input Options**

* Plain prompt: "Generate synthetic hospital admission dataset"
* Structured schema: JSON or CSV header + types + constraints
* Seed dataset (optional): to model conditional distributions

**🧰 Tech Stack**

| **Layer** | **Tools/Libraries** |
| --- | --- |
| Generation | PyTorch, CTGAN, TVAE, SDV |
| Validation | Pandera, Cerberus, custom rule engine |
| Profiling | ydata-profiling, Deepchecks, pandas |
| Agents | Python (OOP), multiprocessing |
| UI/API (optional) | Streamlit / FastAPI |
| Storage | CSV, SQLite, Parquet |

**🌍 Long-Term Vision**

* Extend to time-series and relational synthetic data
* Integrate with ArguBot for intelligent data debate and refinement
* Provide an open-source alternative to commercial synthetic data platforms
* Become a drop-in synthetic data API for ML pipelines

**🔹 Conclusion**

AutoSynth isn’t just about making fake data — it’s about **building intelligence into the data generation process**. With agents, learning, and evolution baked in, it allows engineers, scientists, and testers to finally stop hunting for data and start **creating it smartly**.