College Name: VIT Bhopal
Student Name: Garvit

GEN AI PROJECT PHASE 1 SUBMISSION DOCUMENT

Phase 1: Proposal & Idea Submission

1. Project Title:

Synthetic Soil Data Generation using CTGAN

2. Domain:

Generative AI | Synthetic Data Generation | Tabular Data

3. Problem Statement:

Limited real-world datasets can hinder the development and validation of robust machine learning models, especially in domains like agriculture and soil science. Obtaining and labeling large amounts of real soil data can be expensive and time-consuming. This project addresses this challenge by leveraging generative AI to create synthetic soil data that resembles real data statistically and can be used for model training and analysis.

4. Proposed Solution:

This project will implement a synthetic data generation pipeline using the CTGAN (Conditional Tabular Generative Adversarial Networks) model. The system:

- Takes a real soil dataset as input, containing various soil properties and features.
- Trains the CTGAN model to learn the underlying patterns and distributions of the data.
- Generates synthetic soil data samples that mimic the statistical properties of the original dataset.
- Evaluates the quality and fidelity of the synthetic data using various metrics.

5. Objectives:

- To build a working prototype that generates synthetic soil data using CTGAN.
- To fine-tune the CTGAN model for optimal performance on the specific soil dataset.
- To evaluate the quality of the synthetic data using statistical and visual comparisons.
- To demonstrate the utility of the synthetic data for downstream tasks like machine learning model training.

6. Expected Outcome:

- A synthetic soil dataset with a significant number of samples.
- A comprehensive evaluation report comparing the real and synthetic datasets.
- A trained CTGAN model that can be used to generate more synthetic data on demand.
- A demonstration of how the synthetic data can be used to improve the performance of machine learning models.

7. Tools & Technologies to be Used:

- Python (Primary programming language)
- SDV (Synthetic Data Vault) library
- CTGAN model from SDV
- Pandas and NumPy for data manipulation
- Scikit-learn for model evaluation
- Matplotlib and Seaborn for data visualization

- Jupyter Notebook for experimentation
- Google Colab for model training and data generation

8. References:

- SDV Documentation:
- CTGAN Research Paper: