College Name: VIT Bhopal
Student Name: Garvit

GEN AI PROJECT PHASE 3 SUBMISSION DOCUMENT

Phase 3: Final Report and Submission

1. Project Title:

Synthetic Data Generation for Soil Data using CTGAN

2. Summary of Work Done

Phase 1 – Proposal and Idea Submission:

In this phase, we identified the challenge of limited and potentially sensitive soil data for agricultural research and modeling. We proposed the idea of leveraging Generative AI to create synthetic soil data that mirrors the statistical properties of real soil data while preserving privacy. The objectives were defined as follows:

- Explore the capabilities of CTGAN (Conditional Tabular Generative Adversarial Networks) for synthetic data generation.
- Understand the process of data preprocessing, model training, and evaluation in the context of synthetic data generation.
- Generate synthetic soil data that maintains the statistical distributions and relationships of the original dataset.
- Evaluate the quality and utility of the synthetic data.

We submitted a detailed proposal outlining the problem definition, objectives, tools (Python, CTGAN, Pandas, Scikit-learn), and expected outcomes.

Phase 2 – Execution and Demonstration:

In the second phase, we implemented the proposed idea using Python libraries such as Pandas, Scikit-learn, and the SDV (Synthetic Data Vault) library for CTGAN implementation. The following tasks were accomplished:

- Data Acquisition and Preprocessing: Obtained real soil data and performed necessary preprocessing steps like handling missing values and encoding categorical variables.
- CTGAN Model Training: Trained the CTGAN model using the preprocessed soil data to learn the underlying data distributions.
- Synthetic Data Generation: Generated a synthetic dataset of soil data using the trained CTGAN model.
- Data Evaluation and Visualization: Evaluated the quality of the synthetic data using statistical metrics and visualizations, comparing distributions and relationships with the original data.

College Name: VIT Bhopal
Student Name: Garvit

3. GitHub Repository Link

You can access the complete codebase, README instructions, and any related resources at the following GitHub link: https://github.com/garvit-exe/Synthetic-Soil-Data-Generation-using-CTGAN

4. Testing Phase

4.1 Testing Strategy

The synthetic data generation process was rigorously tested to ensure the quality and utility of the generated data. The testing phase focused on evaluating the statistical properties and relationships within the synthetic data compared to the original dataset.

4.2 Types of Testing Conducted

- **Statistical Similarity Testing:** Compared distributions of individual variables and relationships between variables in the real and synthetic datasets.
- **Visualization:** Used plots and charts to visually assess the similarity between real and synthetic data distributions.
- **Qualitative Assessment:** Examined the synthetic data for realistic values and patterns.

4.3 Results

- **Statistical Similarity:** The synthetic data demonstrated a high degree of similarity to the original data in terms of distributions and relationships.
- **Visualizations:** Plots and charts confirmed the visual similarity between real and synthetic data.
- **Qualitative Assessment:** The synthetic data exhibited realistic values and patterns, indicating the model's ability to capture underlying data characteristics.

5. Future Work

While the project successfully demonstrated the generation of synthetic soil data using CTGAN, there are areas for future enhancement:

- **Model Fine-tuning**
  Explore hyperparameter optimization techniques to further improve the quality of synthetic data.
- **Advanced Evaluation Metrics**
  Investigate the application of more advanced evaluation metrics specific to synthetic data quality.
- **Integration with Downstream Tasks**
  Apply the synthetic data to real-world agricultural tasks, like crop yield prediction or soil management optimization, to assess its practical utility.

College Name: VIT Bhopal
Student Name: Garvit

6. Conclusion

This project successfully demonstrated the potential of **Generative AI**, specifically **CTGAN**, for creating **high-quality synthetic soil data**. By replicating the statistical properties of real data, this approach offers a valuable tool for agricultural research and modeling while addressing privacy concerns associated with sensitive data. Future work will focus on refining the model, exploring advanced evaluation methods, and integrating the synthetic data into practical applications.