DataMirage - A Unified Platform for Synthetic Data Generation

Team Members

Arjan Sapkota (THA077BCT012)

Girban Adhikari (THA077BCT017)

Jivan Acharya (THA077BCT019)

Subarna Ghimire (THA077BCT043)

Supervised By:

Er. Umesh Kanta Ghimire

HOD

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

June 21, 2024

Presentation Outlines

- Motivation
- Objectives
- Scope of Project
- Proposed Methodology
- Expected Results
- Projects Applications
- Gantt Chart
- Estimated Project Expenses
- References

Motivation

- Increasing challenges in leveraging data for AI applications
  - Growing AI model complexity demands larger, high-quality datasets
- Traditional data collection is costly and time-intensive
  - Gathering and processing real-world data requires significant resources
- Ethical and privacy concerns with real data
  - Real data use risks privacy violations and ethical issues

Objectives

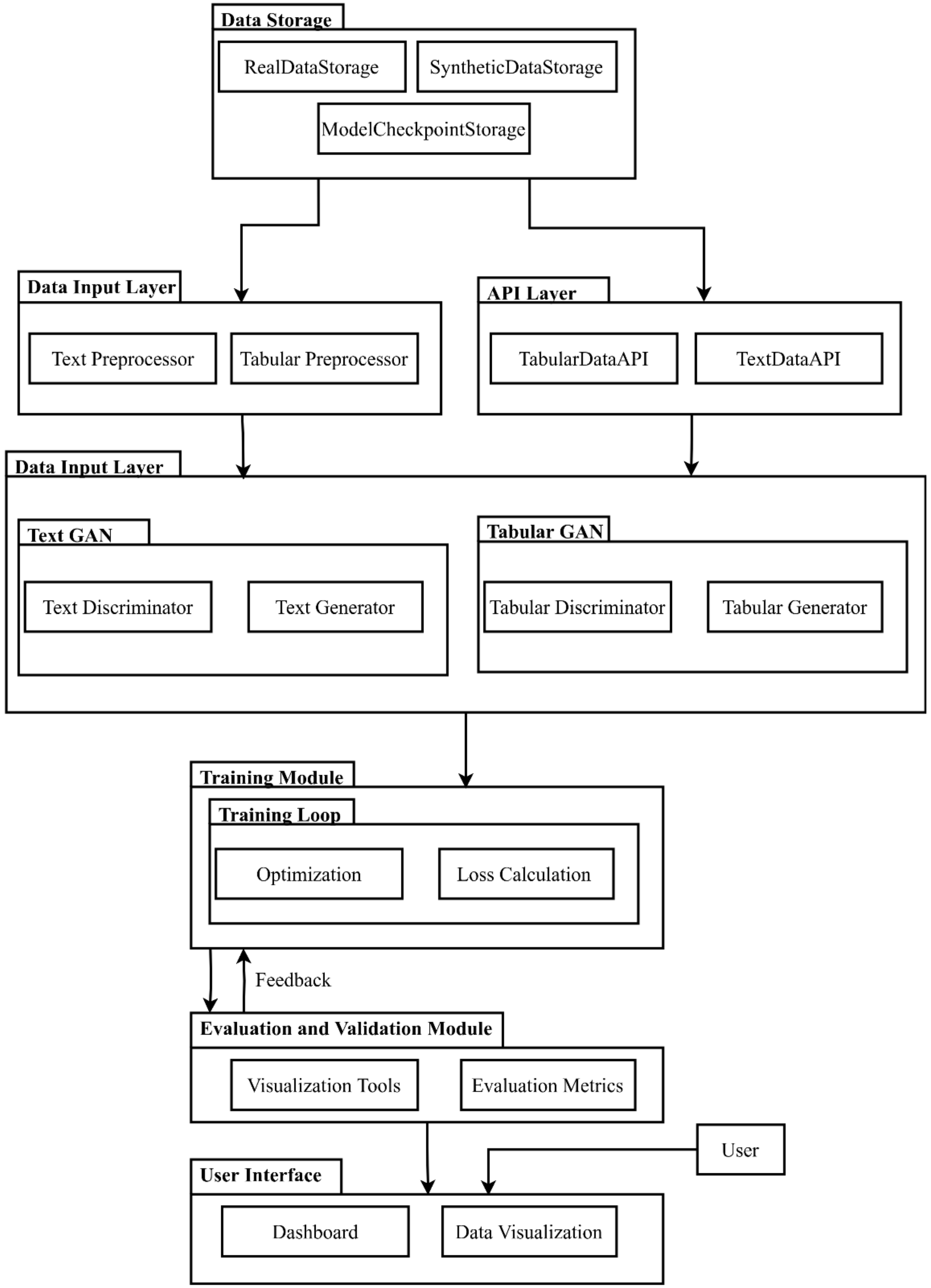
- Develop a platform for generating high-quality synthetic data across tabular and textual datasets

- Enhance machine learning model training with privacy-preserving synthetic data

Scope of Project

- Project Capabilities:
  - Generate diverse synthetic data for various datasets
  - Replace sensitive data to ensure privacy compliance
  - Improve AI model accuracy with augmented synthetic data
- Project Limitations:
  - Synthetic data may lack perfect realism, affecting model performance
  - High-quality generation is computationally intensive and resource-demanding
  - Regulatory bodies may not accept synthetic data for all applications.

Proposed Methodology – [1] (System Implementation Diagram)



Proposed Methodology – [2] (Working Principle)

- Data Storage
  - RealDataStorage: Stores the original datasets that will be used to generate synthetic data
  - SyntheticDataStorage: Contains the synthetic datasets generated by the system
  - ModelCheckpointStorage: Keeps track of model checkpoints for saving progress and continuing training

Proposed Methodology – [3] (Working Principle)

- Data Input Layer
  - Text Preprocessor: Processes and prepares textual data for synthetic data generation
  - Tabular Preprocessor: Processes and prepares tabular data for synthetic data generation

- API Layer
  - TabularDataAPI: Interface for accessing and manipulating tabular data
  - TextDataAPI: Interface for accessing and manipulating textual data

Proposed Methodology – [4] (Working Principle)

- Data Generation Models
  - Text GAN
    - Text Discriminator: Evaluates the quality of generated textual data
    - Text Generator: Produces synthetic textual data that mimics real data
  - Tabular GAN
    - Tabular Discriminator: Assesses the authenticity of generated tabular data
    - Tabular Generator: Generates synthetic tabular data that replicates the real data

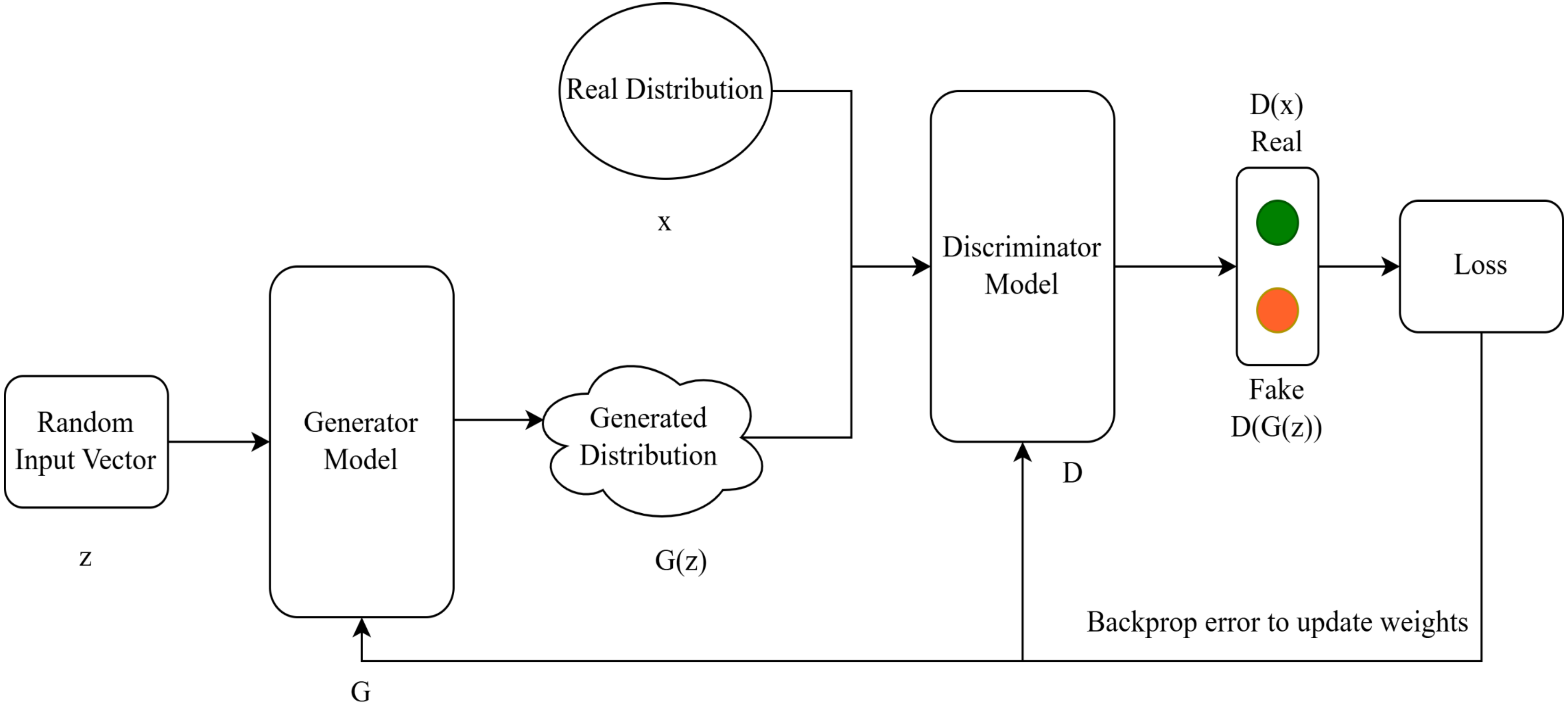
- Training Module
  - Training Loop
    - Optimization: Adjusts model parameters to improve the quality of synthetic data
    - Loss Calculation: Computes the difference between generated data and real data to guide training

Proposed Methodology – [5] (Working Principle)

- Evaluation and Validation Module
  - Visualization Tools: Provides graphical representations to analyze and interpret data
  - Evaluation Metrics: Offers metrics to assess the quality and validity of the synthetic data

- User Interface
  - Dashboard: Central hub for monitoring and managing the data generation process
  - Data Visualization: Tools to visualize both the real and synthetic datasets for better understanding and comparison

Proposed Methodology – [6] (Architecture of GAN)



Proposed Methodology – [7] (Working Principle)

- Basic Structure
  - Generator (G)
    - Takes random noise as input
    - Generates synthetic data resembling real data
  - Discriminator (D)
    - Takes both real and synthetic data as input
    - Outputs the probability that the input data is real

Proposed Methodology – [8] (Working Principle)

- Adversarial Process
  - Training Phase
    - Step 1: Train Discriminator
      - Real data labeled as real
      - Synthetic data from the generator labeled as fake
      - Discriminator learns to distinguish between real and fake data
    - Step 2: Train Generator
      - Generator produces synthetic data
      - Synthetic data is fed to the discriminator
      - Generator learns to produce data that fools the discriminator into classifying it as real

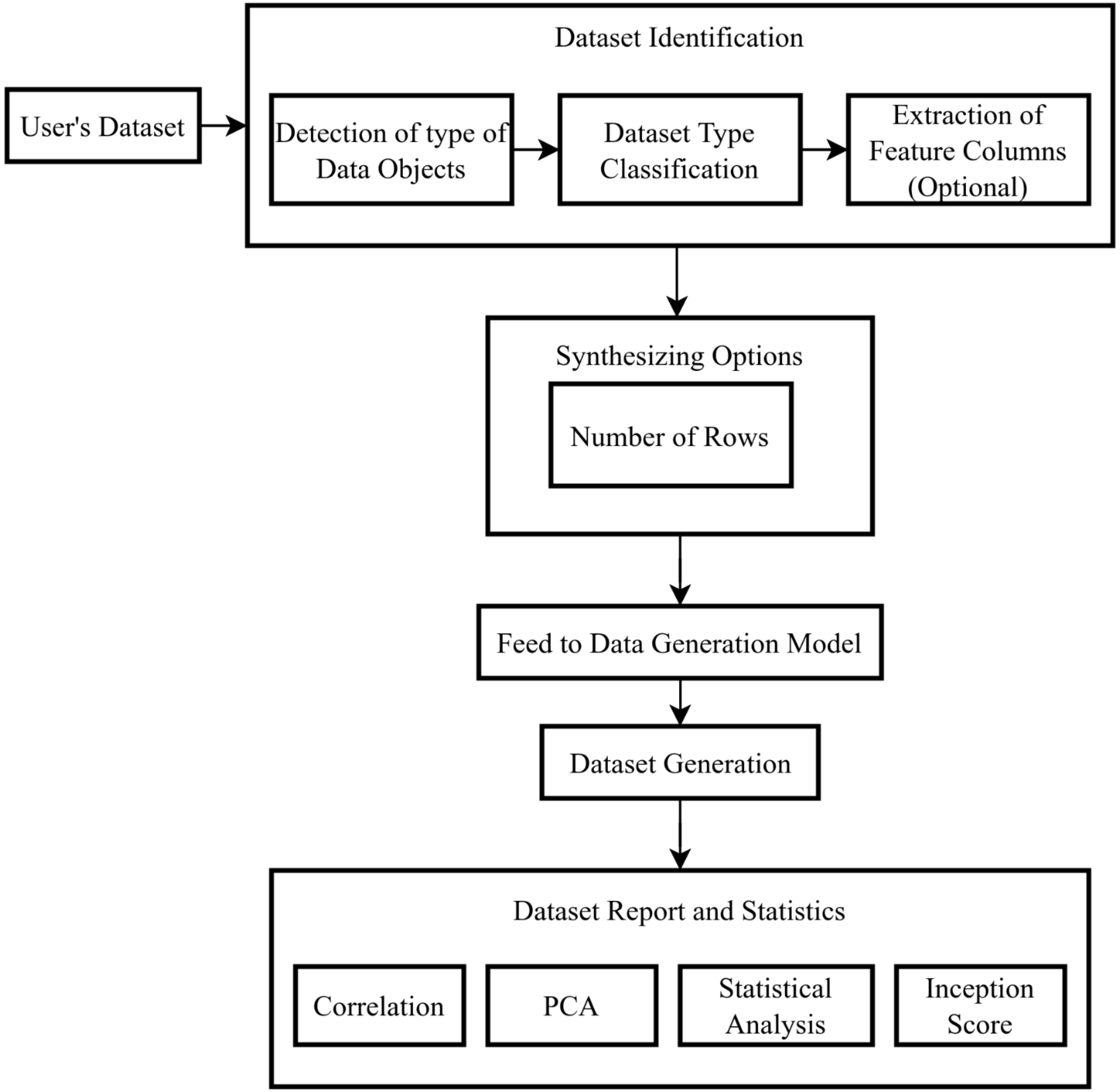
Proposed Methodology – [9] (Working Principle)

- Objective Functions
  - Discriminator Loss
    - Measures the accuracy of the discriminator in distinguishing real data from synthetic data
  - Generator Loss
    - Measures how well the generator can produce data that the discriminator classifies as real
- Iterative Training
  - Alternating optimization steps for the discriminator and generator
  - Discriminator: Maximizes Discriminator loss
  - Generator: Minimizes Generator loss

Proposed Methodology – [10] (Working Principle)

- Convergence
  - The process continues until the generator produces data that the discriminator can no longer distinguish from real data
  - Ideally, both networks reach a point where: D(x)=0.5
    for real and fake data

Proposed Methodology – [11] (System Flow)



Proposed Methodology – [12] (Working Principle)

- Detects the type of data objects in the user's dataset
- Classifies the dataset type and optionally extracts feature columns
- User specifies the number of rows for the synthetic dataset
- Input data and synthesizing options are fed into the data generation model
- System generates synthetic data based on the provided specifications
- Generates a comprehensive report including correlation, statistical analysis to validate

Proposed Methodology – [13] (Hardware Requirements)

- Processor:
  - NVIDIA Tesla K80, P100, or T4 (Google Colab)
  - NVIDIA Tesla P100 (Kaggle)
- RAM:
  - Up to 25 GB (Google Colab)
  - 13 GB (Kaggle)
- Persistent Storage:
  - 5 GB per notebook (Kaggle)
- GPU Access:
  - Free access to powerful GPUs (Google Colab)

Proposed Methodology – [14] (Software Requirements)

- Programming Languages: Python
- Development Environments and IDEs: Jupyter Notebook, Google Colab, Kaggle Kernels
- Data Processing and Analysis: Pandas, NumPy, Scikit-learn
- Deep Learning Frameworks: TensorFlow, Keras, PyTorch
- Synthetic Data Generation: GANs - TensorFlow and PyTorch
- Model Training and Evaluation: TensorBoard, Weights & Biases
- Data Storage and Management: Google Drive, Kaggle Datasets
- Version Control: GitHub

Dataset Exploration – [1] (Textual)

Attribute	Details
Dataset Name	Mental Health Counselling Conversations
Data Type	Textual
Source	Primarily User-Contributed
Size	3.51k rows
Information Covered	
Context	String containing the question asked by a user
Response	String containing the corresponding answer provided by a psychologist

Dataset Exploration – [2] (Textual)

Context string · *lengths* 25⇔293 63.8%	Response string · *lengths* 0⇔3.27k 98.8%
How can I get to a place where I can be content from day to day?	It's important to take a look inside and see what's going on with you to cause you to have these feelings. Please contact us in whatever way is most comfortable for you and we can…
I have a severe back problem. I've had 3 major and several minor operations, but I'm still in constant pain. How can I deal with the depression from this chronic pain?	Chronic pain at the back likely results from a few areas:L4-L5 kidney zone, most likely (lower back);Bone spurs, fused discs, and slipped discs, caused by connective tissue weakness, and calcium deposits used to neutralize highly acidic areas...The 'depression' will evaporate when the chronic pain is drained out, through natural means;Pharmaceutical means will simply extend the pain and cause it to deepen over time, not solving the problem;Remember, medical doctors suppress, natural doctors cure...
I have a severe back problem. I've had 3 major and several minor operations, but I'm still in constant pain. How can I deal with the depression from this chronic pain?	Maybe if you started to address questions of an inner nature of what changed in your life as a result of the back problem.To know your limitations and the areas of your life which…
I suffer from adult ADHD, anxiety disorder, and depression. It has been difficult to find a doctor in my area and my primary physician won't help. I am unemployed and overwhelmed.…	If it is simply counseling that you seek, any number of faith-based outfits are very willing to listen and help out with these sorts of matters, free of charge :)Online…

Dataset Exploration – [3] (Tabular)

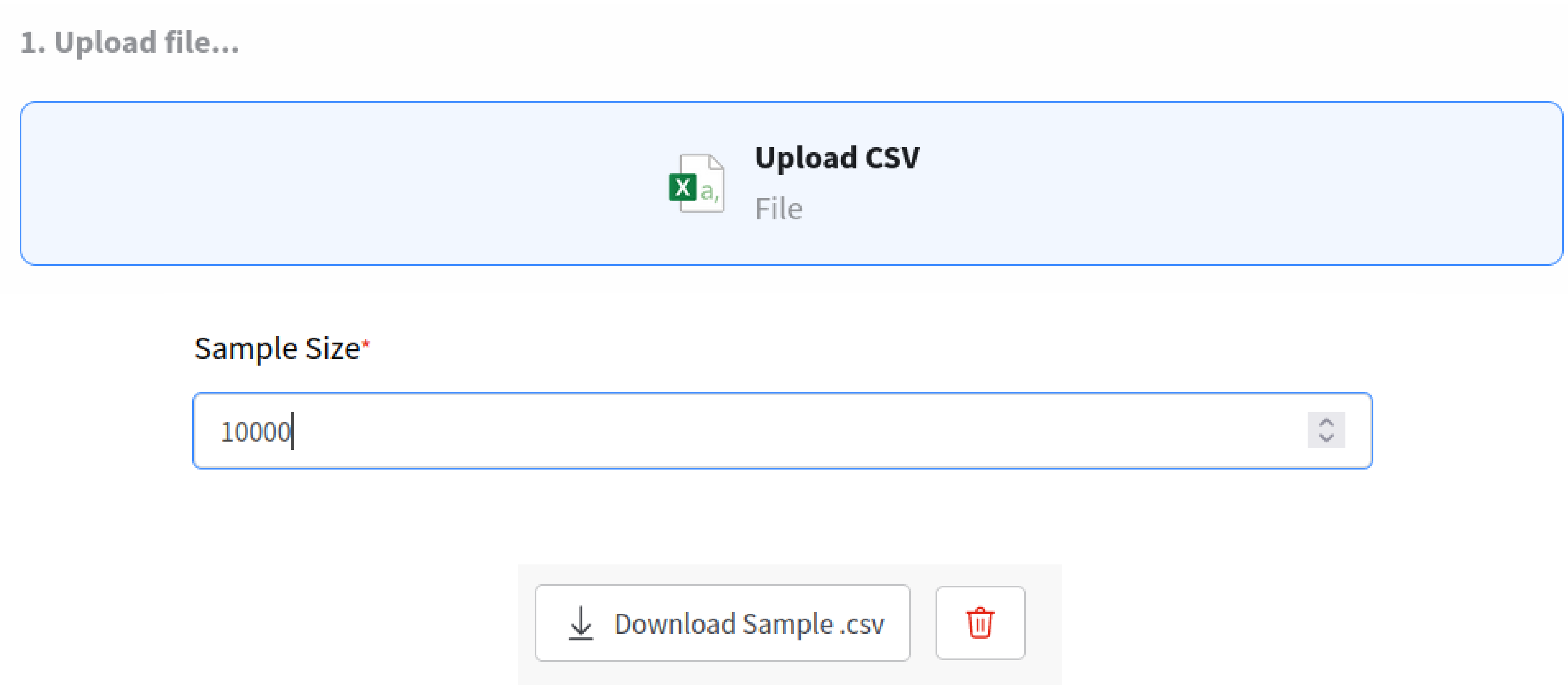
Dataset Name	Data Type	Source	Size (No. of Instances)	Covered Information	Features
Database 1 to 6	Boolean or Continuously-valued	Garavan Institute	~2800 (training, 972 testing)	Various	~29 attributes
Database with 9172 instances	Boolean or Continuously-valued	Ross Quinlan	9172	Covers 20 classes, includes domain theory	~29 attributes
Thyroid database by Stefan Aeberhard	Boolean or Continuously-valued	Stefan Aeberhard	215	Thyroid condition, no missing values	5 attributes
ANN-suited Thyroid database	Boolean or Continuously-valued	Peter Turney	3772 (training, 3428 testing)	Thyroid condition	3 classes, includes cost data

Expected Results – [1]

You don't have Synthetic data created yet.
Synthesizers generate new data from a specific Data Source.

Create Synthetic Data

Expected Results – [2]



Project Applications

- Privacy-Preserving Applications
  - Substituting sensitive data with synthetic equivalents to mitigate privacy risks
  - Enhancing AI model training without compromising sensitive health/financial data
- AI Model Training and Performance
  - Augmenting existing datasets with synthetic data to boost model accuracy
  - Facilitating faster iteration and deployment of AI solutions in various fields
- Educational and Training Purposes
  - Providing realistic synthetic datasets for training researchers, students, and professionals
  - Enabling practical experimentation with accessible and diverse datasets

Gantt Chart


Jun-24
Jul-24
Sep-24
Oct-24
Dec-24
Feb-25
Research
Deployment
Coding
Testing and Debugging
Designing
Documentation and Report

Estimated Project Expenses

TASK	EXPECTED PRICE (NRs)
Printing	2500.00
Compute Resources	10000.00
Deployment	3000.00
Total	**15500.00**

References – [1]

[1] I. J. Goodfellow, J. Pouget-Abadie and M. Mirza, "Generative Adversarial Networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672-2680

[2] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations (ICLR)*, 2013.

[3] E. Choi, S. Biswal and B. Malin, "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks," in *Proceedings of Machine Learning Research*, 2017.

References – [2]

[4] Y. Zhang, Z. Gan and K. Fan, "Adversarial Feature Matching for Text Generation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, August.

[5] L. Yu, W. Zhang and J. Wang, "Sequence generative adversarial nets with policy gradient," in *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[6] R. Quinlan, "Thyroid Disease," UCI Machine Learning Repository, 1987.[Online]. Available: https://doi.org/10.24432/C5D010.