CTGAN
Conditional Tabular Generative Adversarial Network
for synthetic healthcare data generation

Colab link :
https://colab.research.google.com/drive/1cK1aJDKg96NinTkzEBiKUgPSB5VqkysS?usp=sharing

Bhanu Arya
ba27444

Tutorial

1. We first understand the **need for Synthetic Data** in Healthcare.
2. We then look at CTGAN, the **underlying algorithm** used to generate synthetic data
3. We will then look at the **code implementation** of CTGAN. Specifically:
    a. Loading and Pre-process MIMIC data
    b. Train CTGAN model to learn distribution of MIMIC data
    c. Generate similar synthetic data
    d. Evaluating the generated synthetic data

1. Challenges of Healthcare Data - solved by Synthetic Data

Topic	Problem	Solution
Privacy & Collaboration	Healthcare data is sensitive and restricted by regulations like HIPAA.	Synthetic data preserves statistical properties while ensuring privacy, enabling secure collaboration.
Data Availability & Balance	Limited data for rare diseases hinders research and AI training.	Synthetic data augments small datasets and balances representation, improving AI fairness.
AI Training & Testing	AI models can be biased due to imbalanced demographics or conditions.	Techniques like GANs generate realistic data, enabling robust training and edge-case testing.
Education & Simulation	Real patient data is restricted for training professionals and simulating rare events.	Synthetic patient cases provide safe, realistic training scenarios for medical learning and AI readiness.

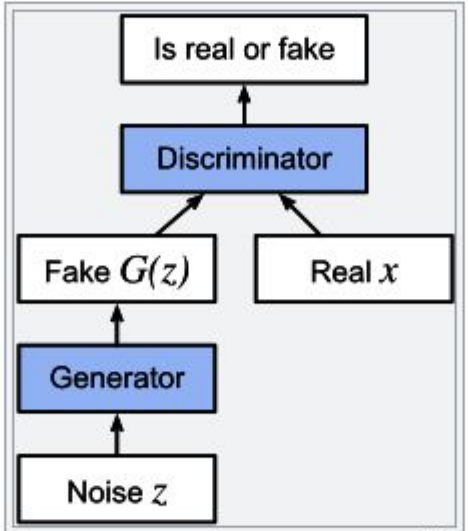
Deep learning algorithms like GAN can learning the distribution of real data and generated synthetic data, overcoming these problems.

2. How Generative Adversarial Network works?

A Generative Adversarial Network (GAN) is a deep learning model consisting of two neural networks:

- Generator – Creates synthetic data (e.g., images) to mimic real data.
- Discriminator – Evaluates whether the data is real or generated

They compete in a zero-sum game, improving each other through training. The generator tries to create realistic outputs, while the discriminator refines its ability to distinguish real from fake. Over time, the generator produces highly realistic data.



An illustration of how a GAN works

2. How CTGAN - works?

CTGAN (Conditional Tabular GAN) is a type of GAN designed to generate realistic synthetic tabular data. It improves on standard GANs by addressing the challenges of learning from tabular datasets with mixed data types (numerical and categorical).

Key features of CTGAN:

- Mode-specific Normalization – Handles skewed numerical data distributions.
- Conditional Generator – Balances imbalanced categorical data by conditioning on specific categories.
- Training with GAN Framework – Uses a generator to synthesize tabular data and a discriminator to distinguish real from fake, refining both through adversarial training.

CTGAN is useful for privacy-preserving data generation, augmenting small datasets, and training ML models without real data.

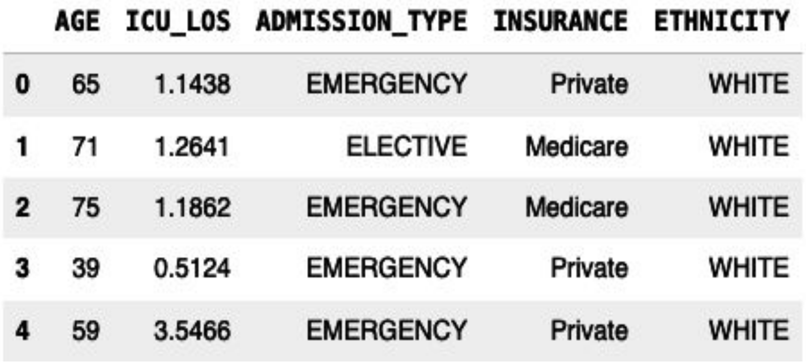
3.1. Code - Load and Preprocess MIMIC-III Data

Loading Data >>

1. Load MIMIC-III Data: Reads compressed CSV files for admissions, patients, and ICU stays.

Data Processing and Feature Engineering >>

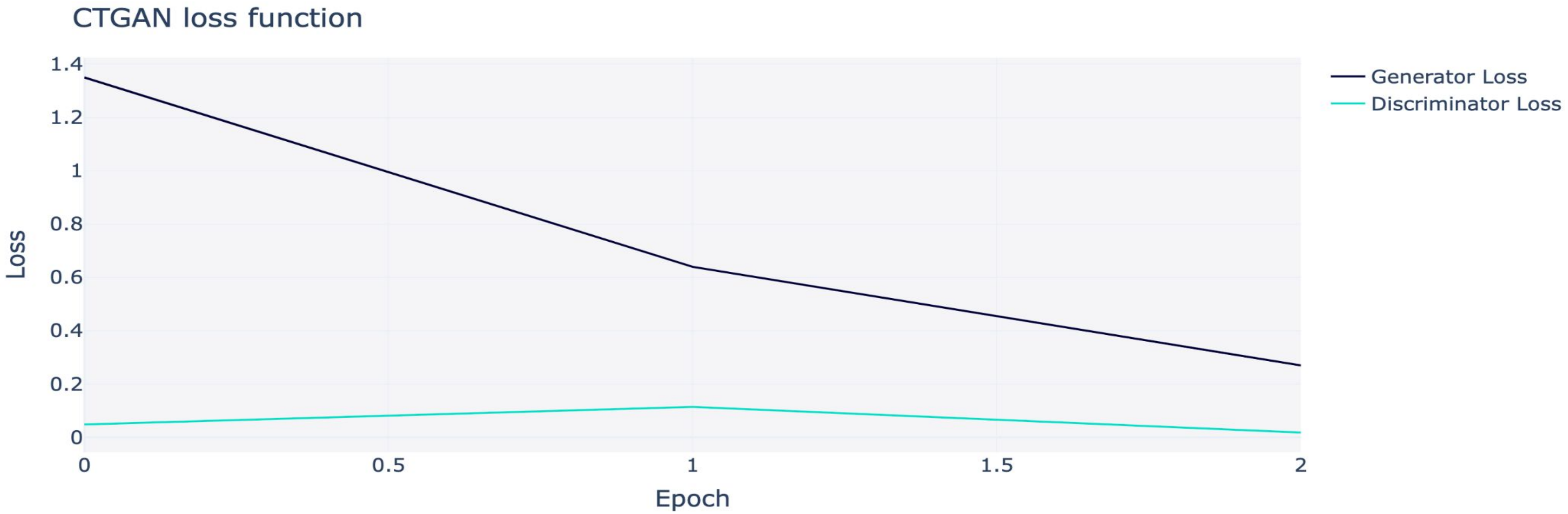
2. Select Relevant Columns: Filters only the necessary columns from each dataset.
3. Convert Date Columns: Transforms `ADMITTIME` and `DOB` to datetime format.
4. Compute Age at Admission: Merges patient data to compute age at admission.
5. Merge ICU Stay Data: Combines ICU length of stay (LOS) with admissions data.
6. Keep Essential Features: Retains only `AGE`, `ICU_LOS`, `ADMISSION_TYPE`, `INSURANCE`, and `ETHNICITY`.
7. Handle Missing Values: Replaces missing categorical values with `"Unknown"` and numerical values with `-1`.
8. Convert Categorical Columns: Ensures categorical features are stored as strings.



	AGE	ICU_LOS	ADMISSION_TYPE	INSURANCE	ETHNICITY
0	65	1.1438	EMERGENCY	Private	WHITE
1	71	1.2641	ELECTIVE	Medicare	WHITE
2	75	1.1862	EMERGENCY	Medicare	WHITE
3	39	0.5124	EMERGENCY	Private	WHITE
4	59	3.5466	EMERGENCY	Private	WHITE

3.2. Code - Train CTGAN on MIMIC-III Data

Training CTGAN model on real MIMIC III data on A100 GPU in colab. We save the model after training for future generation of synthetic data. Once model is trained, it no longer requires real data for generation.



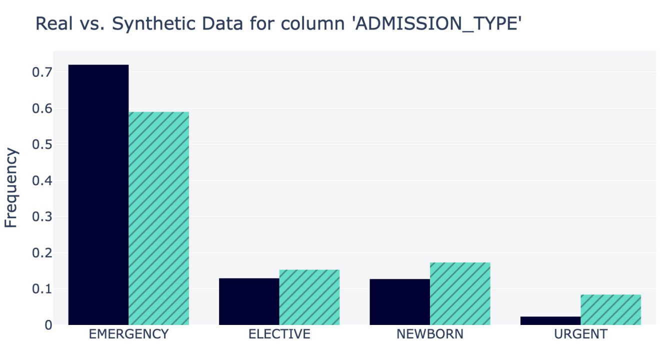
Observe that the Generator and Discriminator loss are converging, even though stopped after 2 epochs (due to limited compute). This means that CTGAN model can train and improve further.

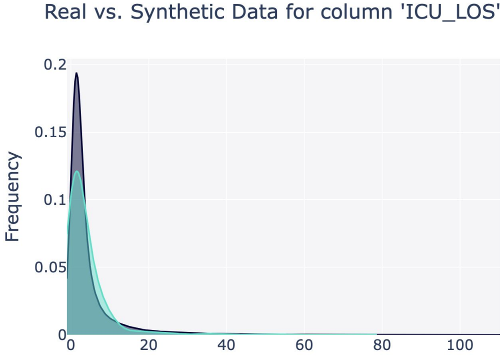
3.3. Code - Generate Synthetic Data

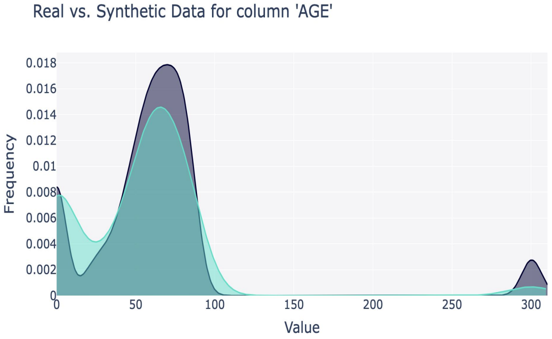
Data

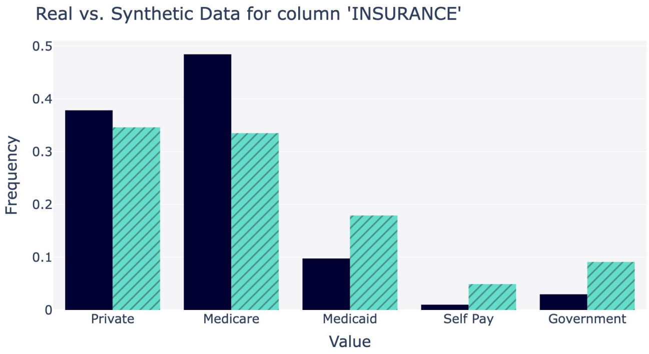
■ Real

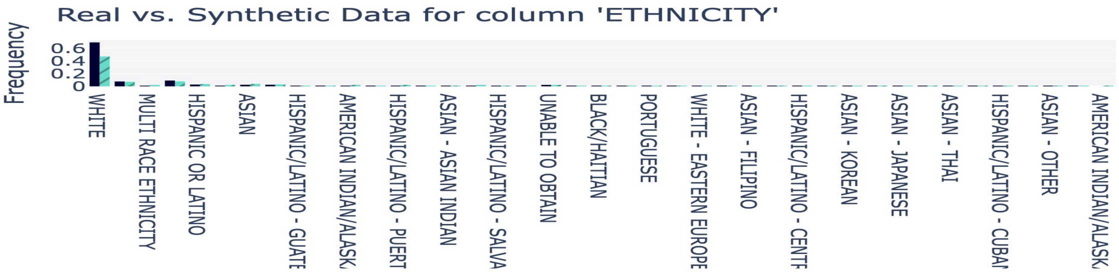
▨ Synthetic


Real vs. Synthetic Data for column 'ADMISSION_TYPE'


Real vs. Synthetic Data for column 'ICU_LOS'


Real vs. Synthetic Data for column 'AGE'


Real vs. Synthetic Data for column 'INSURANCE'

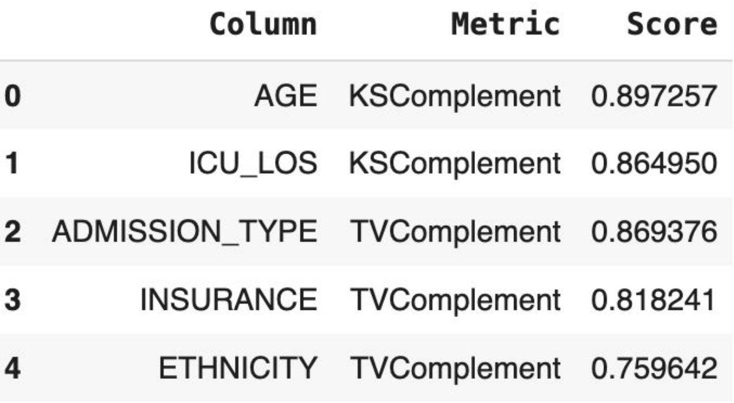

Real vs. Synthetic Data for column 'ETHNICITY'

Observe that high overlap or similar frequency of synthetic and real data distribution after training

3.4. Evaluate Synthetic Data

KS Complement Score : Kolmogorov-Smirnov (KS) test measures the maximum difference between the cumulative distribution functions (CDFs) of real and synthetic data.
- A score close to **1** means the distributions are highly similar, while a lower score indicates a larger divergence.

TV Complement Score (Total Variation Complement Score)
The Total Variation (TV) distance quantifies the difference between two probability distributions.
- A score near **1** indicates that the synthetic data closely matches the real data distribution.



	Column	Metric	Score
0	AGE	KSComplement	0.897257
1	ICU_LOS	KSComplement	0.864950
2	ADMISSION_TYPE	TVComplement	0.869376
3	INSURANCE	TVComplement	0.818241
4	ETHNICITY	TVComplement	0.759642

The high score ranges .75 to .89, indicating the Synthetic data has similar distribution with real data.

Key Takeaways:

CTGAN is GAN based model for conditional data generation on tabular data.

Once CTGAN model is training it can effectively generate synthetic data with similar distribution to MIMIC III.

This enables privacy and safe AI model training, overcoming some challenges of data in healthcare domain.

Reference:

- https://paperswithcode.com/paper/modeling-tabular-data-using-conditional-gan
- https://arxiv.org/pdf/1907.00503