Data Synthetics Project Proposal

Domain Background

Applicable Domains: Data Synthetics can be applied to almost all domains. But in this project will be applied to **Healthcare** domain.

Historical Information: The idea of original fully synthetic data was created by **Donald Bruce Rubin**. Rubin originally designed this to synthesize the Decennial Census long-form responses for the short-form households to preserve their privacy. and it can be applied to **Healthcare** domain.

What's Data-Synthetics? Synthetic data is information that's artificially generated rather than produced by real-world events.

Common use cases: Privacy concerned domains such as ***Healthcare Domain.*** Also used to generated more data to **improve model accuracy**.

Problem Statement

What's the problem?. When you visit the hospital for a health issue or regular check, you always give the hospital important information about yourself *e.g.,* your diet, medicines you use, health concerns, sensitive data about yourself .etc.

But you want to preserve your privacy and don't want your information out? At the same time, some scientists need this information for to help future patients to come up with a cure or something to help others! and of course, scientists want to preserve your privacy.

Solution Statement

Why Synthetic Data in Healthcare?. Synthetic data can protect patient privacy and augment clinical research as synthetic data is a derivative of the original real data but no synthetic data point can be attributed to a single real data point.

Datasets and Inputs

The dataset is a tabular data that contains 60 rows of blood analysis of Rheumatoid arthritis patience. The data set was obtained from a hospital in Cairo, Egypt.

Data is approved only for the context of this project, and all ids can't be traced to this dataset.

Benchmark Model

The best-known model in Tabular Data Generation is Causal-TGAN (Wen et al, 2021) with a KS test average score of 0.81 for adult, census, and news datasets.

Evaluation Metrics

KS test metrics will be used for evaluation, which is: **Kolmogorov–Smirnov test** which compares the distribution of one or two samples.

Project Design

This project should be approached as follows:

1- Analyize variables distributions as the generative model should be able to generate the same data distribution.

2- Picking a generative model which could be one of the next options:

- GANs model
- AutoEncoder
- Boltzmann Machine
- Diffusion Model
- Using existing package for tabular data generation

3- Make evaluation to model the output

4- Visualizing generated data distribution

Platform to use:

- Will use AWS SageMaker service for training and deploying the model.
- Will use AWS Lambda for processing data triggered by uploading files to s3 buckets.