IMPLEMENTASI GENERASI DATA SINTETIS BERBASIS CTGAN UNTUK PRESERVASI PRIVASI PADA DATASET TELCO CUSTOMER CHURN

Dosen Pengampu Mata Kuliah: Ir. Sidik Prabowo, S.T., M.T., CEH, IDPP

Universitas Telkom

Disusun Oleh:

Muhammad Karov Ardava Barus	103052300001
Muhammad Al Fayyedh Denof	103052330042
Avatar Bintang Ramadhan	103052300007
Runa Raditya Rizki Hidayat	103052300037

PROGRAM STUDI S1 SAINS DATA
FAKULTAS INFORMATIKA
UNIVERSITAS TELKOM
BANDUNG

2025

Daftar Isi

I. Pendahuluan

1.1 Latar Belakang

Dalam era transformasi digital, data pelanggan menjadi aset strategis bagi perusahaan telekomunikasi untuk memahami perilaku konsumen dan mencegah perpindahan pelanggan (*churn*). Dataset *Telco Customer Churn* memuat berbagai atribut sensitif, mulai dari demografi hingga pola penggunaan layanan. Dalam siklus hidup data (*Data Lifecycle Management*), penggunaan data riil (*production data*) untuk keperluan pengembangan model *Machine Learning* (ML) atau pengujian sistem seringkali menimbulkan risiko keamanan yang signifikan.

Risiko utama yang dihadapi adalah serangan *re-identification*, di mana penyerang dapat menggabungkan data yang telah dianonimisasi dengan sumber informasi eksternal untuk mengungkap identitas individu. Praktik penggunaan data asli tanpa perlindungan yang memadai tidak hanya membahayakan privasi pelanggan tetapi juga berpotensi melanggar regulasi perlindungan data yang berlaku. Oleh karena itu, diperlukan pendekatan alternatif yang dapat menyediakan data berkualitas tinggi untuk analisis tanpa mengekspos informasi sensitif individu.

1.2 Tujuan

Tujuan utama dari tugas besar ini adalah:

1. Mengimplementasikan *pipeline* generasi data sintetis yang aman untuk dataset *Telco Customer Churn*.
2. Menghasilkan data tiruan yang mematuhi prinsip-prinsip perlindungan privasi sesuai dengan Undang-Undang Perlindungan Data Pribadi (UU PDP) No. 27/2022.
3. Memastikan data sintetis yang dihasilkan tetap mempertahankan utilitas statistik dan korelasi antar fitur sehingga layak digunakan sebagai pengganti data asli dalam pelatihan model prediksi *churn*.

II. Metodologi

2.1 Pendekatan *Synthetic Data*

Untuk mengatasi tantangan privasi tersebut, kami mengajukan penggunaan pendekatan *Synthetic Data Generation* (Opsi 1). Data sintetis adalah data yang dibuat secara artifisial yang meniru karakteristik statistik dari data asli tanpa memuat informasi yang dapat diidentifikasi secara langsung dari subjek data yang sebenarnya.

2.2 Alat dan Algoritma

Kami akan memanfaatkan pustaka *open-source* SDV (Synthetic Data Vault), dengan fokus utama pada penggunaan algoritma CTGAN (Conditional Tabular GAN).

CTGAN dipilih karena keunggulannya dalam menangani data tabular yang kompleks, yang terdiri dari campuran kolom numerik (kontinu) dan kategorikal (diskrit). Berbeda dengan metode statistik tradisional atau GAN standar, CTGAN menggunakan *mode-specific normalization* untuk mengatasi distribusi data yang rumit dan *conditional generator* untuk menangani

ketidakseimbangan kategori (*imbalanced data*), yang sangat relevan dengan karakteristik dataset *churn*.

2.3 Alur Kerja (*Pipeline*)

Proses pengerjaan akan mengikuti tahapan berikut:

1. Preprocessing: Membersihkan data asli dan melakukan *encoding* yang diperlukan.
2. Model Training: Melatih model CTGAN menggunakan data asli untuk mempelajari distribusi probabilitas gabungan dari fitur-fitur data.
3. Sampling: Membangkitkan sampel data baru (sintetis) dari model yang telah dilatih.
4. Post-processing: Memastikan format data sintetis sesuai dengan struktur data asli.

III. Rencana Evaluasi

Keberhasilan pembentukan data sintetis akan dievaluasi berdasarkan dua dimensi utama:

3.1 Evaluasi Privasi

Kami akan mengukur risiko privasi untuk memastikan data sintetis aman dari serangan inferensi. Metrik yang akan digunakan meliputi:

- K-Anonymity: Memastikan bahwa setiap rekaman dalam data tidak dapat dibedakan dari setidaknya $k - 1$ rekaman lainnya.
- Distance to Closest Record (DCR): Mengukur jarak Euclidean antara data sintetis dan data asli untuk memastikan tidak ada data sintetis yang merupakan duplikat persis atau terlalu mirip dengan data asli (*overfitting*).

3.2 Evaluasi Utilitas

Kami akan mengukur seberapa baik data sintetis mempertahankan informasi dari data asli:

- Utilitas Statistik: Membandingkan distribusi variabel tunggal (histogram/KDE) dan korelasi antar variabel (*correlation matrix*) antara data asli dan sintetis.
- Machine Learning Efficacy: Menggunakan metode *Train on Synthetic, Test on Real* (TSTR). Kami akan melatih model klasifikasi (seperti Random Forest atau XGBoost) menggunakan data sintetis dan menguji akurasinya pada data asli. Penurunan performa yang minim menandakan utilitas data yang tinggi.