DESAIN PIPELINE GENERASI DATA SINTETIS BERBASIS CTGAN

Progress Project Keamanan Data – Week 3

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

Disusun Oleh:

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

PROGRAM STUDI S1 SAINS DATA
FAKULTAS INFORMATIKA
UNIVERSITAS TELKOM BANDUNG

2025

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi data dalam industri telekomunikasi membuat perusahaan semakin bergantung pada kemampuan analisis data pelanggan untuk memahami pola penggunaan layanan dan memprediksi risiko perpindahan pelanggan atau churn. Namun, meningkatnya kebutuhan analisis ini diiringi pula dengan meningkatnya tantangan dalam menjaga privasi data. Dataset Telco Customer Churn yang digunakan dalam project ini mengandung berbagai atribut sensitif yang berpotensi menimbulkan risiko pelanggaran privasi apabila digunakan secara langsung, terutama dalam konteks pengembangan model Machine Learning.

Regulasi Undang-Undang Perlindungan Data Pribadi (UU PDP) No. 27/2022 menjadi dasar hukum yang mempertegas pentingnya proses pengolahan data yang aman dan sesuai prinsip data minimization serta pseudonimisasi. Oleh karena itu, penggunaan data sintetis menjadi solusi yang semakin diminati karena dapat menghadirkan alternatif data yang menyerupai data asli tanpa mengekspos identitas individu. Pada tahap Week 3 ini, kelompok kami memfokuskan diri pada perancangan pipeline generasi data sintetis menggunakan algoritma CTGAN (Conditional Tabular GAN), yang merupakan teknologi generatif berbasis deep learning yang dirancang khusus untuk data tabular dengan kombinasi fitur kategorikal dan numerik.

Pipeline yang dirancang pada minggu ini menjadi fondasi utama untuk implementasi pada minggu-minggu berikutnya. Oleh sebab itu, desain pipeline perlu dijelaskan dengan detail agar memudahkan proses coding, debugging, validasi, dan evaluasi performa model di tahap selanjutnya.

1.2 Tujuan

Tahap Week 3 tidak berfokus pada proses implementasi langsung, melainkan pada penyusunan konsep pipeline yang matang dan terdokumentasi dengan baik. Tujuan dari tahap ini adalah menghasilkan rancangan menyeluruh mengenai bagaimana data sintetis akan dibangun, apa saja komponen yang terlibat, dan bagaimana keterkaitan antar setiap prosesnya. Dengan adanya blueprint yang jelas, proses implementasi di Week 4 akan lebih terarah.

BAB II
DESAIN PIPELINE GENERASI DATA SINTESIS

2.1 Tahap Preprocessing Data

Tahap pertama yang dilakukan adalah persiapan dataset asli Telco Customer Churn. Dataset ini mengandung beberapa fitur yang memiliki tipe data berbeda, sehingga perlu melalui proses pembersihan dan normalisasi terlebih dahulu sebelum digunakan untuk melatih model CTGAN.

Pada tahap ini, penanganan missing values menjadi langkah awal untuk mencegah bias dan ketidaksesuaian pada proses training. Selain itu, beberapa kolom yang tidak relevan atau mengandung informasi identitas langsung perlu dipertimbangkan untuk dihapus atau dijadikan acuan pseudonimisasi. Fitur kategorikal yang berbentuk teks kemudian dikonversi menjadi representasi numerik menggunakan teknik encoding agar dapat dikenali oleh model deep learning.

Kolom numerik juga dievaluasi untuk menentukan apakah perlu dilakukan normalisasi atau transformasi tertentu, terutama jika terdapat rentang nilai yang terlalu jauh. Keseluruhan proses preprocessing ini bertujuan untuk menghasilkan dataset yang bersih, konsisten, dan siap digunakan pada tahap pelatihan model CTGAN.

2.2 Tahap Training Model CTGAN

Setelah data diproses, tahap berikutnya adalah melatih model CTGAN menggunakan dataset tersebut. CTGAN merupakan model berbasis arsitektur GAN yang terdiri dari dua komponen utama, yaitu generator dan discriminator. Generator bertugas menghasilkan data baru, sementara discriminator berusaha membedakan mana data asli dan mana data sintetis.

CTGAN memiliki keunggulan dalam menangani data tabular karena menggunakan mekanisme mode-specific normalization untuk fitur numerik dan conditional generation untuk fitur kategorikal. Dengan mekanisme ini, CTGAN dapat mengatasi tantangan seperti distribusi data yang tidak seimbang (imbalanced data) serta fitur kategorikal yang memiliki banyak kelas.

Proses pelatihan dilakukan secara iteratif hingga generator mampu memproduksi data yang sulit dibedakan dari data asli oleh discriminator. Pada tahap ini, parameter seperti jumlah epoch, batch size, serta learning rate menjadi hal yang perlu diperhatikan untuk mendapatkan model yang stabil.

2.3 Tahap Generasi (Sampling) Data Sintesis

Setelah model berhasil dilatih, model kemudian digunakan untuk menghasilkan sampel data sintetis. Proses ini disebut sebagai *sampling*, yaitu mengambil data dari ruang distribusi probabilistik yang telah dipelajari oleh generator.

Jumlah data sintetis yang dihasilkan direncanakan minimal sebanyak 1.500 baris sesuai ketentuan pada KAK. Meski demikian, jumlah sampel dapat disesuaikan dengan kebutuhan evaluasi yang akan dilakukan pada minggu berikutnya. Pada tahap ini, perlu dilakukan pemeriksaan visual dan statistik awal untuk memastikan bahwa data sintetis yang dihasilkan memiliki distribusi yang mendekati data asli tanpa meniru secara langsung.

2.4 Tahap Post-processing Data Sintetis

Tahap terakhir dalam pipeline adalah proses post-processing. Pada tahap ini, data sintetis yang dihasilkan dikembalikan ke dalam format yang sama dengan dataset asli agar dapat digunakan pada proses evaluasi model Machine Learning.

Jika pada tahap preprocessing dilakukan encoding, maka pada tahap ini encoding tersebut harus dibalik kembali menjadi label kategorikal yang sesuai. Selain itu, beberapa kolom numerik yang seharusnya berupa bilangan bulat seperti tenure perlu dibulatkan kembali. Konsistensi format dan tipe data perlu diperhatikan agar dataset sintetis terlihat seperti data tabular normal.

Tahap ini juga mencakup pemeriksaan final terhadap apakah ada duplikasi atau kemiripan ekstrem dengan data asli yang dapat berpotensi menimbulkan risiko privasi.

BAB III
TEKNIK YANG DIPILIH: CTGAN DAN ALASAN PEMILIHAN

3.1 Karakteristik CTGAN

CTGAN merupakan model GAN yang dirancang secara khusus untuk bekerja dengan data tabular, berbeda dari GAN konvensional yang biasanya digunakan pada data gambar atau sinyal kontinu. Salah satu karakteristik paling penting dari CTGAN adalah penggunaan *conditional generator*, yaitu mekanisme yang memungkinkan model untuk belajar dari distribusi setiap kolom secara lebih terarah. Fitur ini sangat membantu terutama ketika data memiliki ketidakseimbangan kelas, karena model dapat menghasilkan sampel baru yang tetap memperhatikan proporsi kategori tertentu. Selain itu, CTGAN juga mampu menghasilkan data sintetis berkualitas tinggi meskipun dataset awal memiliki struktur yang kompleks, banyak variabel, atau distribusi yang tidak mengikuti pola umum.

3.2 Kelebihan CTGAN

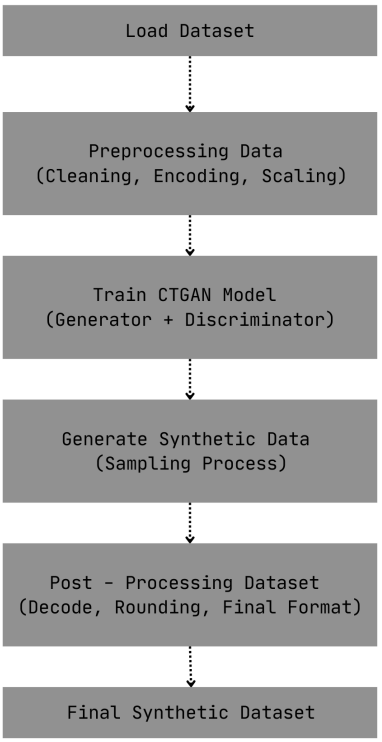
Dibandingkan dengan pendekatan pembangkitan data lainnya, CTGAN memiliki sejumlah kelebihan yang membuatnya lebih layak digunakan dalam proyek sintesis data tabular. Salah satu keunggulan utamanya adalah kemampuannya menangani variabel kategorikal dan numerik secara natural tanpa memerlukan proses encoding manual yang terlalu rumit. Ini sangat relevan pada dataset besar yang memiliki banyak atribut campuran. CTGAN juga bekerja lebih baik daripada GAN konvensional dalam menjaga struktur distribusi campuran tersebut, sehingga data sintetis yang dihasilkan lebih stabil dan lebih mirip dengan pola statistik data asli. Selain itu, CTGAN merupakan metode yang direkomendasikan oleh SDV (Synthetic Data Vault), yang telah menjadi standar dalam banyak penelitian terkait data sintetis. Fakta bahwa CTGAN digunakan luas di berbagai publikasi makin memperkuat kredibilitas dan keandalannya.

3.3 Kesesuaian dengan Dataset Telco Customer Churn

Jika ditinjau dari karakteristik dataset Telco Customer Churn, CTGAN menjadi teknik yang sangat sesuai. Dataset churn memiliki banyak kolom kategorikal seperti jenis layanan, metode pembayaran, atau tipe kontrak; sekaligus banyak kolom numerik seperti tenure dan besaran tagihan bulanan. Kombinasi ini membutuhkan model yang mampu memahami dua tipe data sekaligus. Selain itu, dataset churn biasanya memiliki *class imbalance* karena jumlah pelanggan yang churn jauh lebih sedikit dibandingkan yang tidak churn. CTGAN, dengan conditional generator-nya, dapat mengatasi ketimpangan tersebut dengan lebih efektif. Dataset ini juga memiliki beberapa fitur sensitif dan *quasi-identifiers*, sehingga model yang digunakan harus mampu menjaga privasi tanpa meniru data asli secara terlalu dekat. CTGAN memenuhi kebutuhan ini karena mampu mempertahankan hubungan antar kolom, menghasilkan data yang tetap realistik, dan secara bersamaan meminimalkan risiko re-identification pada individu.

BAB IV
FLOWCHART PIPELINE GENERASI DATA SINTESIS

4.1 Diagram Flowchart



Flowchart generasi data sintetis dimulai dari dataset asli Telco Customer Churn, yang dipersiapkan melalui pre-processing seperti pembersihan data, penanganan nilai hilang, dan transformasi fitur. Data yang telah siap kemudian digunakan untuk melatih CTGAN, yang mempelajari distribusi dan hubungan antar fitur, termasuk menangani ketidakseimbangan kelas melalui conditional generation.

Setelah model terlatih, dilakukan sampling untuk menghasilkan data sintetis yang meniru pola data asli tanpa menyalin identitas pengguna. Hasil sintetis tersebut melalui post-processing agar format dan nilainya konsisten. Terakhir, data diuji melalui evaluasi privasi dan utilitas, sebelum menjadi output sintetis yang aman dan tetap representatif untuk analisis atau pelatihan model churn.

BAB V
PENUTUP

5.1 Output

Pada tahap Week 3, kelompok kami berhasil menyelesaikan beberapa komponen persiapan penting, yaitu penyusunan pipeline generasi data sintetis yang lengkap dan terstruktur, penjabaran setiap tahap dalam bentuk naratif, serta pembuatan diagram flowchart yang memudahkan pemahaman alur kerja. Seluruh rancangan ini akan menjadi dasar dalam pengembangan model dan evaluasi privasi-utilitas pada minggu berikutnya.

5.2 Kesimpulan

Berdasarkan karakteristik, kelebihan, dan kesesuaiannya dengan struktur dataset Telco Customer Churn, CTGAN dapat disimpulkan sebagai teknik yang paling tepat untuk digunakan dalam proses generasi data sintetis pada proyek ini. Model ini tidak hanya mampu menangani kombinasi variabel kategorikal dan numerik secara efektif, tetapi juga menawarkan mekanisme *conditional generation* yang secara signifikan membantu dalam mengatasi ketidakseimbangan kelas — salah satu tantangan umum pada dataset churn. Kualitas data yang dihasilkan CTGAN juga terbukti stabil dan realistis, sehingga secara statistik tetap mencerminkan pola penting yang dimiliki data asli.

Di sisi lain, CTGAN tetap menjaga aspek privasi melalui kemampuannya menghasilkan data baru yang tidak identik dengan data asli namun tetap memiliki utilitas yang tinggi. Hal ini menjadi nilai tambah yang sangat penting dalam konteks perlindungan data pelanggan. Dengan seluruh pertimbangan tersebut, pemilihan CTGAN tidak hanya didasarkan pada keunggulan teknis, tetapi juga pada kesesuaian metodologinya terhadap kebutuhan analisis churn yang kompleks dan sensitif. Oleh karena itu, CTGAN menjadi fondasi utama dalam pipeline generasi data sintetis yang dibangun pada proyek ini.