ANALISIS RISIKO PRIVASI DAN VISUALISASI 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, sebelum dilakukan proses sintesis data atau pemodelan lebih lanjut, diperlukan analisis mendalam mengenai karakteristik data dan risiko privasi yang terkandung di dalamnya.

1.2 Tujuan

Tujuan dari laporan tahap ini adalah:

1. Mengidentifikasi variabel sensitif (*Direct Identifiers*, *Quasi-Identifiers*, dan *Sensitive Attributes*) pada dataset *Telco Customer Churn*.
2. Menganalisis tingkat kerentanan privasi data menggunakan metrik *K-Anonymity* untuk mendeteksi risiko *re-identification*.
3. Memvisualisasikan distribusi data awal untuk memahami karakteristik dan pola *churn*.

II. Identifikasi Variabel Data

Dataset *Telco Customer Churn* terdiri dari 7.043 baris data pelanggan dengan 21 atribut. Langkah pertama dalam menjaga privasi data adalah mengklasifikasikan setiap atribut berdasarkan tingkat sensitivitas dan risikonya.

Berikut adalah deskripsi atribut dalam dataset:

Nama Kolom	Deskripsi	Tipe
customerID	ID unik pelanggan	Direct ID
gender	Jenis kelamin pelanggan (Male/Female)	QI
SeniorCitizen	Apakah pelanggan warga senior (1/0)	QI
Partner	Apakah memiliki pasangan (Yes/No)	QI
Dependents	Apakah memiliki tanggungan (Yes/No)	QI
tenure	Lama berlangganan (bulan)	Numerik
PhoneService	Layanan telepon (Yes/No)	Kategorikal
MultipleLines	Banyak saluran telepon	Kategorikal
InternetService	Penyedia layanan internet (DSL/Fiber/No)	Kategorikal
OnlineSecurity	Layanan keamanan online	Kategorikal
OnlineBackup	Layanan backup online	Kategorikal
DeviceProtection	Layanan perlindungan perangkat	Kategorikal
TechSupport	Layanan dukungan teknis	Kategorikal
StreamingTV	Layanan streaming TV	Kategorikal
StreamingMovies	Layanan streaming film	Kategorikal
Contract	Jangka waktu kontrak	Sensitif
PaperlessBilling	Tagihan tanpa kertas	Kategorikal
PaymentMethod	Metode pembayaran	Kategorikal
MonthlyCharges	Biaya bulanan	Sensitif
TotalCharges	Total biaya	Sensitif
Churn	Status berhenti berlanggan (Yes/No)	Target/Sensitif

Table 1: Deskripsi Atribut Dataset Telco Customer Churn

2.1 Direct Identifiers (Pengenal Langsung)

Atribut ini dapat secara langsung menunjuk pada satu individu tertentu tanpa perlu informasi tambahan.

- Daftar Atribut: `customerID`
- Tindakan: Atribut ini wajib dihapus atau dilakukan proses *hashing* (pseudonimisasi) sebelum data digunakan untuk pemodelan, karena memiliki risiko re-identifikasi 100%.

2.2 Quasi-Identifiers (QI)

Atribut ini sendiri mungkin tidak unik, namun jika dikombinasikan dengan atribut lain (atau data eksternal), dapat digunakan untuk mengidentifikasi individu (*linkage attack*).

- Daftar Atribut (Sampel): `gender`, `SeniorCitizen`, `Partner`, `Dependents`
- Analisis: Kombinasi dari atribut-atribut demografis ini akan digunakan untuk menghitung skor *K-Anonymity*. Semakin unik kombinasi nilai-nilai ini, semakin tinggi risiko privasinya.

2.3 Sensitive Attributes (Atribut Sensitif)

Informasi rahasia yang menjadi nilai intrinsik dari dataset dan harus dilindungi kerahasiaannya terhadap subjek data.

- Daftar Atribut: `Churn` (Target), `Contract`, `MonthlyCharges`, `TotalCharges`
- Tujuan: Memastikan bahwa meskipun seseorang berhasil diidentifikasi (yang harus dicegah), penyerang tetap tidak boleh mengetahui nilai atribut sensitif ini secara pasti (*l-diversity*).

III. Analisis Risiko Privasi

Kami menggunakan metrik K-Anonymity untuk mengukur risiko re-identifikasi.

3.1 Landasan Teori K-Anonymity

K-Anonymity adalah konsep privasi yang menjamin bahwa setiap individu dalam dataset tidak dapat dibedakan dari setidaknya $k-1$ individu lain berdasarkan himpunan atribut *Quasi-Identifier* (QI).

Definisi Matematis: Misalkan T adalah tabel dengan atribut $A_1, ..., A_n$. Misalkan $Q_T \subseteq \{A_1, ..., A_n\}$ adalah himpunan *Quasi-Identifiers*. Tabel T memenuhi k-anonymity jika untuk setiap baris $t \in T$, terdapat setidaknya $k-1$ baris lain $t' \in T$ sedemikian sehingga:

$$\forall A \in Q_T, t[A] = t'[A]$$

Artinya, ukuran setiap kelas ekuivalensi (kelompok baris dengan nilai QI yang sama) harus $\geq k$.

Pseudocode Implementasi (Python): Berikut adalah logika algoritma untuk menghitung nilai k pada dataset:

```python
def calculate_k_anonymity(dataset, quasi_identifiers):
    """
    Menghitung nilai k-anonymity dari dataset.
    """
    # 1. Kelompokkan data berdasarkan kombinasi nilai QI
    #    Setiap grup mewakili satu 'equivalence class'
    grouped_data = dataset.groupby(quasi_identifiers)

    # 2. Hitung ukuran (jumlah baris) setiap grup
    group_sizes = grouped_data.size()

    # 3. Nilai k adalah ukuran grup terkecil yang ditemukan
    k_value = group_sizes.min()

    return k_value
```

3.2 Hasil Perhitungan

Berdasarkan eksekusi kode analisis pada dataset, kami melakukan pengelompokan data berdasarkan QI: `gender`, `SeniorCitizen`, `Partner`, dan `Dependents`.

- Nilai K-Anonymity (k): 3
- Status Risiko: AMAN (Minimal k=3)

Analisis: Nilai $k = 3$ menunjukkan bahwa untuk setiap kombinasi karakteristik demografis yang ada dalam dataset, terdapat minimal 3 orang yang memilikinya. Tidak ada individu yang unik sendirian (k=1). Meskipun statusnya aman, nilai ini masih tergolong rendah, sehingga penerapan teknik privasi tambahan seperti generasi data sintetis tetap direkomendasikan untuk meningkatkan keamanan data sebelum dipublikasikan.

IV. Visualisasi Data Awal

Untuk memahami karakteristik data, kami melakukan visualisasi distribusi data pada atribut target dan demografi.

4.1 Distribusi Target (Churn)

Visualisasi *bar chart* menunjukkan adanya ketidakseimbangan kelas (*class imbalance*) yang signifikan. Jumlah pelanggan yang tidak *churn* (No) jauh lebih dominan (sekitar 5000+) dibandingkan yang *churn* (Yes) (sekitar 1800+). Hal ini mengindikasikan perlunya penanganan khusus saat pelatihan model nantinya agar tidak bias ke kelas mayoritas.

4.2 Distribusi Demografi

Plot distribusi `gender` memperlihatkan proporsi yang seimbang antara pelanggan laki-laki dan perempuan. Selain itu, pola *churn* terlihat serupa pada kedua gender, tanpa perbedaan signifikan.

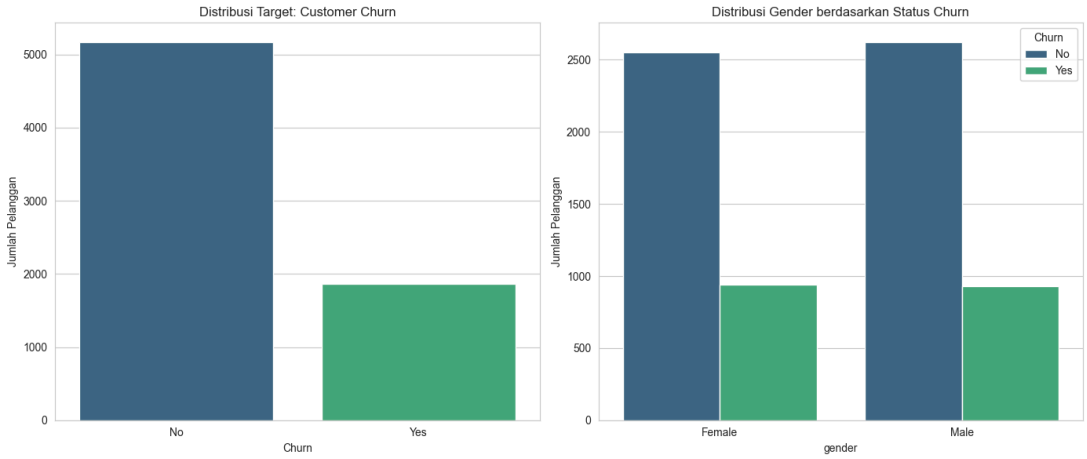

Figure 1: Visualisasi Distribusi Awal Data (Churn & Gender)

V. Kesimpulan

Berdasarkan analisis yang telah dilakukan:

1. Dataset mengandung satu *Direct Identifier* (`customerID`) yang harus dihapus.
2. Analisis risiko privasi menghasilkan nilai K-Anonymity = 3, yang berarti data relatif aman dari serangan re-identifikasi langsung, namun masih memiliki risiko jika dilakukan serangan yang lebih canggih.
3. Visualisasi data menunjukkan adanya ketidakseimbangan pada target `Churn`, yang menjadi catatan penting untuk tahapan pemrosesan data selanjutnya.