Konsolidasi Data Pelanggan dari Banyak Workbook Excel

Studi kasus portofolio - Cerdas Inovasi Digital

Data sintetis, bukan data klien. Semua nama, nomor telepon (pola fiktif) dan email (domain example.*) dibuat untuk studi kasus ini. Angka akurasi di bawah berlaku pada data sintetis ini dan bukan janji hasil pada data nyata; data nyata biasanya lebih beragam.

Masalah

Kontak pelanggan dan riwayat pesanan tersebar di 7 workbook (11 lembar) dengan nama kolom berbeda-beda, telepon dan tanggal dalam format campur, email yang salah ketik, serta pelanggan yang sama muncul di beberapa berkas dengan ejaan nama berbeda. Tugasnya: satu workbook induk tanpa duplikat, kontak **dan** riwayat pesanan tiap pelanggan tetap utuh, ada log perubahan, dan kasus yang tidak pasti tidak digabung diam-diam.

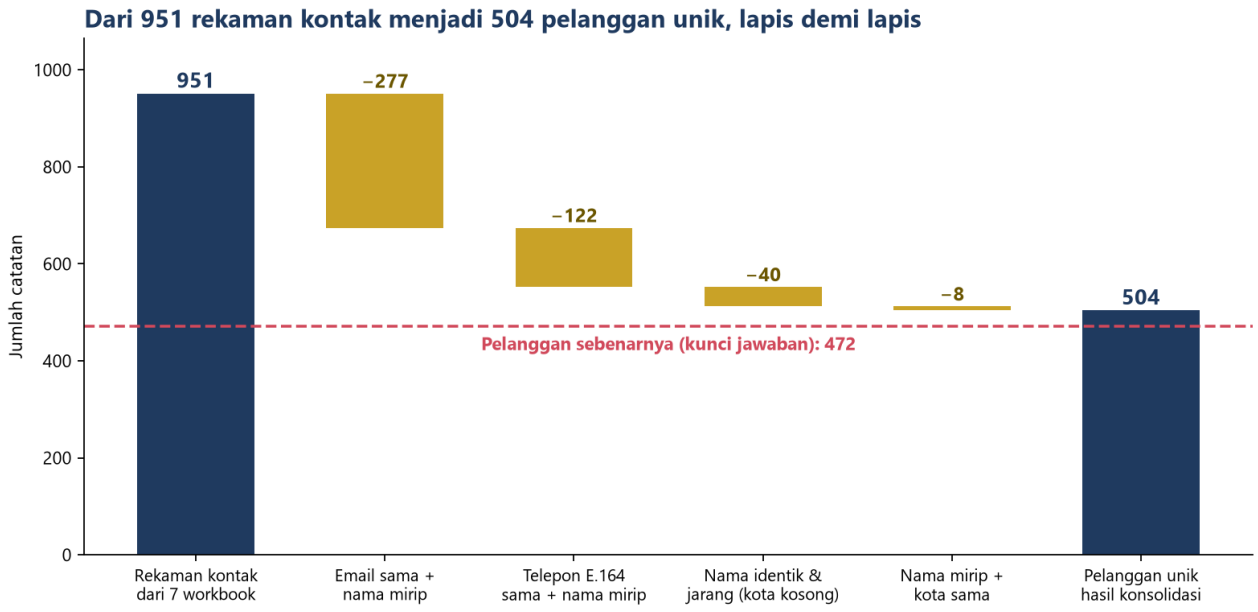
Hasil dalam angka

Ukuran	Nilai
Workbook sumber / lembar data	7 / 11
Rekaman kontak sebelum pencocokan	951
Pelanggan unik sesudah konsolidasi	504 (447 rekaman duplikat digabung)
Pelanggan yang muncul di 2 berkas atau lebih	229
Baris pesanan terbaca / pesanan akhir	1.489 / 1.473 (16 baris ganda persis dibuang)
Total nilai pesanan pada riwayat	Rp 301.795.000
Pasangan di lembar 'Perlu Ditinjau' (tidak digabung)	67

Cara kerja pencocokan berlapis

Lapis	Aturan	Penggabungan
1. Email	Email dinormalkan (huruf kecil, spasi dibuang, domain salah ketik diperbaiki berbasis domain yang paling sering muncul). Sama persis dan nama semirip ≥ 0,80.	277
2. Telepon	Nomor diubah ke E.164 (+62...). Sama persis dan nama semirip ≥ 0,80.	122
3. Nama + kota	Nama dinormalkan (gelar/sapaan dibuang, ejaan lama seperti oe→u, dj→j, ch→h disamakan), kemiripan ≥ 0,93, kota sama, email dan telepon tidak sama-sama bertentangan.	8
3b. Nama identik	Nama ≥ 0,97 dengan 2 kata atau lebih, nama itu dimiliki ≤ 2 pelanggan, kota tak ada pembanding, tak ada kontak bertentangan.	40
Ragu	Email/telepon sama tetapi nama beda, nama sama tetapi kota beda, nama satu kata, nama agak mirip (0,88-0,93): **tidak digabung**, masuk 'Perlu Ditinjau' beserta alasannya.	67

Per pelanggan disimpan rekaman terlengkap (nama dengan dukungan terbanyak dan paling lengkap, email/telepon utama), email dan telepon lain tetap tersimpan sebagai kolom alternatif, kolom kosong diisi dari rekaman duplikatnya, dan seluruh pesanan dari semua berkas dipertahankan.

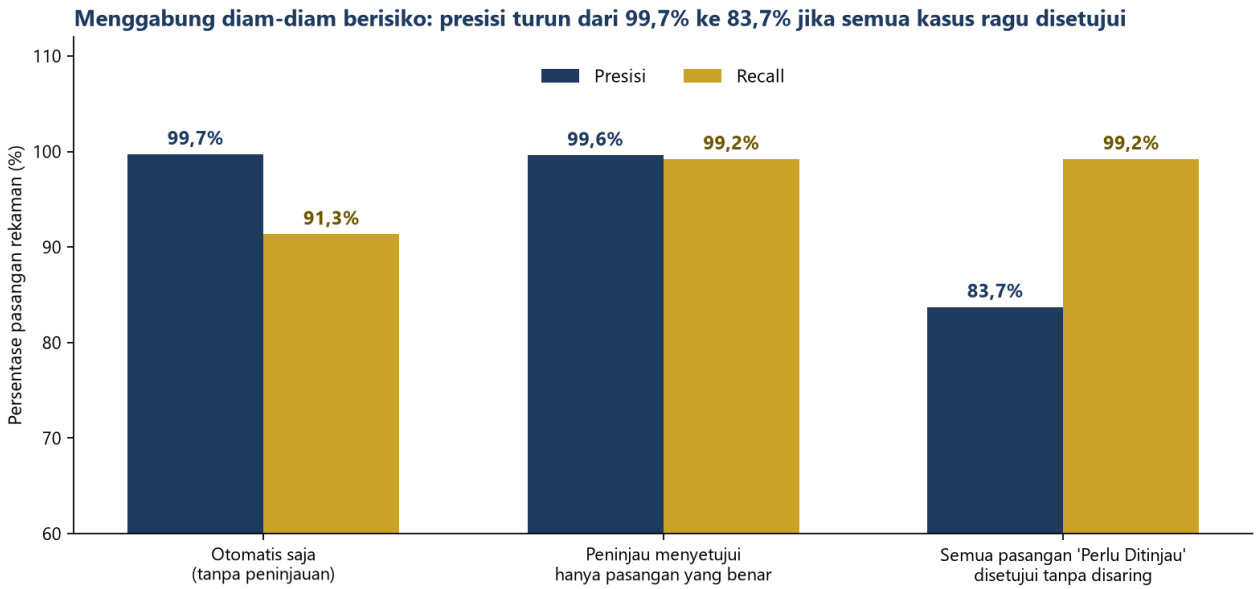


Data sintetis (bukan data klien). Garis merah dari kunci jawaban, dipakai hanya untuk evaluasi akhir.

Evaluasi terhadap kunci jawaban (hanya di akhir)

Kunci jawaban berisi ID pelanggan asli untuk setiap baris sumber dan baru dibuka setelah aturan ditetapkan; ambang tidak disetel ulang menurut kunci. Penilaian dilakukan pada pasangan rekaman: pasangan benar bila kedua rekaman milik pelanggan asli yang sama.

Ukuran	Hasil
Presisi penggabungan otomatis	99,7% (696 dari 698 pasangan yang digabung benar)
Recall penggabungan otomatis	91,3% (696 dari 762 pasangan sejati ditemukan)
F1	95,3%
Pelanggan: dihasilkan vs sebenarnya	504 vs 472
Pelanggan sebenarnya yang masih terpecah / klaster campuran	29 / 1
Pasangan di 'Perlu Ditinjau' yang sebenarnya orang sama	34 dari 67
Pesanan: akhir vs sebenarnya (hilang / ganda tersisa)	1.473 vs 1.473 (0 / 0)
Pesanan menempel pada pelanggan yang benar	99,9%



Data sintetis. Sumbu y dimulai dari 60% agar selisih terlihat. Pasangan dihitung antar-rekaman terhadap kunci ID pelanggan asli.

Yang tersisa dan batasannya

Penggabungan berlebih. Ada 2 pasangan salah gabung (satu klaster campur): dua anggota keluarga yang berbagi kontak dan hampir semua kata di nama (mis. 'Maya Dwi Permata' dan 'Setiyo Dwi Permata') lolos ambang 0,80. Email/telepon bersama tidak cukup sebagai bukti tunggal.

Penggabungan kurang. 66 pasangan sejati belum digabung (29 pelanggan masih terpecah). Pada 34 dari 38 pasangan klaster yang terpecah, kasusnya sudah muncul di 'Perlu Ditinjau'; sisanya tidak punya bukti kontak, kota, atau nama yang cukup. Bila peninjau memilih tepat, recall naik ke 99,2% dengan presisi 99,6%; bila semua pasangan disetujui tanpa disaring, presisi turun ke 83,7%.

Asumsi. Tanggal dd/mm/yyyy dibaca hari-dulu (konvensi Indonesia); 157 tanggal punya hari dan bulan ≤ 12 sehingga bisa ambigu, dicatat di Log Perubahan. Nomor yang disamarkan atau tak valid dikosongkan, tidak ditebak. Produk dinormalkan hanya pada huruf besar/kecil dan satuan 'gr'.

Keterbatasan data sintetis. Pola kesalahan dibuat penulis, jadi aturan yang cocok dengannya belum tentu cocok pada data nyata (misalnya nomor bersama kantor, nama dengan banyak pengguna, atau berkas dengan struktur lain). Di proyek nyata ambang dikalibrasi ulang dengan sampel yang diberi label, dan lembar 'Perlu Ditinjau' ditelaah pemilik data sebelum induk dipakai.

Isi workbook: **Ringkasan**, **Pelanggan** (satu baris per pelanggan, telepon E.164), **Riwayat Pesanan** (tiap pesanan terhubung ke ID pelanggan beserta berkas/baris sumbernya), **Perlu Ditinjau** (pasangan tak pasti dan alasannya), **Log Perubahan** (tindakan pembersihan beserta jumlah dan contoh).