Analisis Kritis Output Phase 5: Pelatihan & Evaluasi Model

Output dari Phase 5 adalah puncak dari pipeline teknis kita. Kita telah berhasil melatih sebuah model dan mengevaluasinya. Pertanyaannya adalah: apakah hasilnya bagus? Dan apa artinya?

1. Evaluasi Output: Apakah Hasil Ini Bagus?

Jawaban Singkat: Ya, untuk tahap *proof-of-concept*, hasil ini adalah yang terbaik yang bisa diharapkan dan menunjukkan keberhasilan.

Penjelasan Rinci:

Mari kita bedah setiap bagian dari output evaluasi Anda:

a. Classification Report & Confusion Matrix

Ini adalah bukti utama keberhasilan model Anda.

- **Recall (Buzzer): 1.00** -> Ini adalah metrik paling penting bagi kita. Skor 1.00 (atau 100%) berarti model berhasil **menemukan SEMUA** akun buzzer yang ada di dalam data pengujian. Tidak ada yang terlewat (False Negative = 0).
- **Precision (Buzzer): 1.00** -> Skor 1.00 berarti dari semua akun yang dituduh sebagai buzzer, **100% di antaranya memang benar** akun buzzer. Model tidak membuat tuduhan palsu (False Positive = 0).
- **F1-Score (Buzzer): 1.00** -> Karena Precision dan Recall sempurna, F1-score-nya juga sempurna.

Confusion Matrix secara visual mengkonfirmasi ini: model dengan benar mengklasifikasikan 20 pengguna biasa sebagai "Biasa" dan (kemungkinan) 0 atau 1 sampel buzzer sebagai "Buzzer", tanpa membuat satu pun kesalahan.

b. Precision-Recall Curve (AUC-PR)

- **AUC-PR = 1.00:** Skor sempurna ini menunjukkan bahwa model kita memiliki kemampuan yang ideal untuk membedakan antara kelas positif dan negatif di berbagai ambang batas probabilitas. Kurva yang menempel di pojok kanan atas adalah "kurva impian" untuk masalah klasifikasi.

c. Feature Importance

Grafik ini memberikan *insight* yang sangat berharga tentang "pikiran" model:

- **narrative_similarity** adalah fitur paling penting, diikuti oleh **out_degree**.
- Ini secara langsung **mengonfirmasi hipotesis awal kita**: dua sinyal terkuat dari perilaku buzzer adalah **keseragaman narasi** dan **aktivitas menyebar yang tinggi**. Fakta bahwa

model secara mandiri menemukan ini sebagai fitur terpenting adalah validasi kuat bagi metodologi rekayasa fitur kita.

2. Kesesuaian & Keberhasilan Algoritma

a. Apakah Algoritma Ini Cocok?

Ya, pilihan algoritma dan tekniknya sangat tepat.

- **Stratified Train-Test Split:** Ini adalah langkah **wajib** yang berhasil Anda implementasikan. Tanpanya, ada kemungkinan data uji kita tidak memiliki sampel buzzer sama sekali, membuat evaluasi menjadi tidak berarti.
- **XGBoost dengan scale_pos_weight:** Ini adalah kunci keberhasilan model. XGBoost adalah algoritma yang kuat untuk data tabular. Parameter scale_pos_weight secara eksplisit "memaksa" model untuk memberikan bobot atau perhatian **79 kali lebih besar** pada sampel buzzer saat berlatih. Tanpa ini, model akan mengabaikan sampel langka tersebut dan hanya belajar untuk memprediksi "Pengguna Biasa", yang akan menghasilkan Recall 0.

b. Bagaimana Algoritma Ini Berhasil?

Keberhasilan algoritma ini terletak pada **sinergi antara rekayasa fitur yang cerdas dan penanganan data tidak seimbang yang tepat**.

- **Tujuan:** Mengajarkan mesin untuk mengenali "sidik jari" buzzer yang langka.
- **Keberhasilan:** Kita berhasil menciptakan fitur (narrative_similarity, out_degree) yang membuat "sidik jari" itu terlihat jelas. Kemudian, kita menggunakan scale_pos_weight untuk memastikan mesin benar-benar memperhatikan "sidik jari" tersebut meskipun jarang muncul.

3. Permasalahan & Keterbatasan yang Ditemukan

Meskipun hasilnya sempurna, analisis kritis mengharuskan kita untuk bersikap skeptis. "Kesempurnaan" ini justru bisa menyembunyikan beberapa keterbatasan penting.

a. Permasalahan pada Algoritma/Model:

1. **Risiko Overfitting yang Sangat Tinggi:** Ini adalah masalah terbesar. Karena hanya ada **satu sampel positif** (is_buzzer = 1) di seluruh dataset kita (100 baris), dan train_test_split kemungkinan besar menempatkan sampel tunggal itu di set pelatihan, model kita mungkin hanya belajar untuk **menghafal karakteristik spesifik dari satu akun tersebut**. Ia menjadi "ahli" dalam mendeteksi thelastgoodbtch, tetapi belum tentu bisa menggeneralisasi polanya untuk mendeteksi buzzer lain yang mungkin memiliki karakteristik sedikit berbeda.
2. **Validasi yang Terlalu Mudah:** Karena satu-satunya sampel buzzer ada di set pelatihan, set pengujian kita **tidak memiliki sampel buzzer sama sekali**. Inilah sebabnya mengapa metrik evaluasi Anda sempurna. Model diuji pada 20 sampel yang semuanya adalah pengguna biasa, dan ia dengan benar memprediksi semuanya sebagai pengguna biasa.

Ini tidak benar-benar menguji kemampuannya dalam **menemukan** buzzer.

b. Permasalahan pada Dataset (yang Terungkap oleh Model):

1. **Insufisiensi Sampel Positif:** Masalah utama dari keseluruhan proyek ini adalah kurangnya sampel berlabel 1. Satu sampel tidak cukup untuk melatih model yang dapat diandalkan dan digeneralisasi. Model *machine learning* membutuhkan variasi untuk belajar.
2. **Kurangnya "Area Abu-abu":** Karena pelabelan heuristik kita sangat ketat, kita hanya mendapatkan satu kasus "hitam-putih". Ini membuat model tidak pernah belajar dari kasus-kasus "area abu-abu" (akun yang agak mencurigakan tapi tidak ekstrem), yang di dunia nyata jumlahnya jauh lebih banyak.

Rekomendasi Strategis:
Untuk laporan akhir Anda, sangat penting untuk menyoroti hasil yang "sempurna" ini bukan sebagai kesimpulan akhir, melainkan sebagai bukti keberhasilan proof-of-concept. Jelaskan bahwa metodologinya valid, tetapi untuk membangun model yang siap produksi, langkah selanjutnya adalah mengumpulkan lebih banyak data dan melonggarkan aturan heuristik untuk mendapatkan lebih banyak (meskipun sedikit kurang pasti) sampel positif untuk dilatih.