**Preprocessing AIS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Nama Tahap** | **Deskripsi** | **Cara Kerja / Kriteria Teknis** | **Manfaat** |
| 1 | **Validasi Skema & Kelengkapan** | Memastikan kolom wajib tersedia dan bertipe benar | Kolom minimal: mmsi, dt\_pos\_utc, latitude, longitude, sog harus ada dan bertipe benar. Jika tidak → file ditolak; jika tipe salah → cast; jika gagal → baris di-drop | Menjamin struktur dan tipe data sebelum diproses lebih lanjut |
| 2 | **Filter Wilayah** | Membatasi cakupan data hanya di wilayah pelabuhan atau laut relevan | Gunakan bounding polygon berbasis H3 resolusi 5 (H3-5) dan filter titik berdasarkan ketercakupan dalam area tersebut | Fokus hanya pada wilayah analisis; mengurangi noise dari luar area |
| 3 | **Deduplikasi Primary Key** | Menghapus duplikat berdasarkan kunci utama data | Primary key = (mmsi, dt\_pos\_utc[, message\_type]); jika ada duplikat, satu disimpan, lainnya di-log atau dihapus | Mencegah kelebihan hitung dalam analisis kapal atau lintasan |
| 4 | **Imputasi Metadata Identitas** | Melengkapi kolom IMO, vessel\_name, dan atribut statis lainnya yang kosong | Imputasi dari data referensi (misalnya: HS database) berdasarkan kecocokan MMSI | Mengurangi missing value pada metadata kapal yang penting untuk analisis spasial dan klasifikasi |
| 5 | **Validasi Konsistensi Identitas** | Pastikan satu identitas kapal = satu set metadata unik | Hitung nunique() dari imo, vessel\_name, length, width per mmsi. Jika ada lebih dari 1 → flag untuk audit | Mendeteksi kasus kesalahan merge atau perubahan nama tak wajar |
| 6 | **Fuzzy Match Nama Kapal (Opsional)** | Memverifikasi konsistensi nama kapal jika satu IMO punya beberapa vessel\_name | Gunakan fuzzy matching (rasio kemiripan teks) untuk menentukan apakah nama berbeda mengacu ke kapal yang sama | Meningkatkan kualitas metadata, menghindari redundansi kapal |
| 7 | **Filter Domain & Outlier Eksplisit** | Buang data yang tidak logis atau tidak relevan | - MMSI 9 digit valid - Draught ≠ 0 - Tipe kapal sesuai (cargo, tanker, dll) - Timestamp dalam periode analisis (mis. tahun 2022) | Membersihkan noise eksplisit dan filter ke kapal yang relevan saja |
| 8 | **Pengujian Rentang Fisik & Format** | Pastikan nilai berada dalam rentang logis | - Latitude: –90° hingga +90° - Longitude: –180° hingga +180° - SOG: 0–70 knots - Waktu: tidak melebihi UTC now + 5 menit | Menyaring titik yang berasal dari error sensor, spoofing, atau data korup |
| 9 | **Koherensi Spasio-Temporal** | Deteksi anomali pergerakan kapal & jeda waktu | - Lompatan >30 km/menit → drop titik - Δt antar ping > 60 menit → flag gap - Nav. status 'Moored'/'Anchored' ↔ SOG harus < 0.5 kn | Mendeteksi error yang sulit dilihat secara kasat mata dan menjaga kualitas trajektori |
| 10 | **Enrichment & Spatial Join** | Tambah informasi kapal & konteks lokasi | - Join dengan data statis (IMO, tipe, panjang) - Spatial join dengan poligon pelabuhan → tandai inside\_port | Menyediakan konteks spasial penting seperti aktivitas pelabuhan, jenis kapal, dan muatan |
| 11 | **Hapus Duplikat Setelah Imputasi** | Pastikan tidak ada duplikat akibat merging data | Jika ada duplikat (mmsi, dt\_pos\_utc) setelah imputasi metadata → drop selektif berdasarkan kualitas data | Menjaga integritas data setelah proses enrikhmen/imputasi otomatis |

**QA AIS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Nama Indikator** | **Deskripsi** | **Cara Hitung** | **Manfaat** |
| 1 | **Missing MMSI** | Persentase baris tanpa informasi identitas kapal (MMSI) | (jumlah mmsi kosong / total baris) × 100 | Menjamin setiap titik bisa dihubungkan ke kapal tertentu |
| 2 | **Missing Waktu Posisi** | Persentase baris tanpa timestamp dt\_pos\_utc | (jumlah waktu kosong / total baris) × 100 | Menentukan urutan dan durasi pergerakan kapal |
| 3 | **Missing Koordinat** | Persentase baris tanpa latitude atau longitude | (jumlah lat/lon kosong / total baris) × 100 | Koordinat adalah inti dari pelacakan posisi |
| 4 | **Missing Kecepatan (SOG)** | Persentase baris tanpa kecepatan | (jumlah sog kosong / total baris) × 100 | Diperlukan untuk analisis manuver dan trafik laut |
| 5 | **MMSI Tidak Valid** | Persentase MMSI yang di luar 9-digit resmi | MMSI < 100.000.000 atau > 999.999.999 | Menyaring noise atau spoofing |
| 6 | **Draughdt = 0** | Persentase baris dengan nilai sarat air nol | (jumlah draught = 0) / total × 100 | Draught 0 artinya sensor tidak aktif atau tak terisi |
| 7 | **Jumlah Rekor** | Total jumlah baris dalam dataset | jumlah total baris | Mengukur cakupan keseluruhan data |
| 8 | **Jumlah Kapal Unik (MMSI)** | Banyaknya kapal unik berdasarkan MMSI | count distinct(mmsi) | Mengidentifikasi jumlah entitas kapal yang tercakup |
| 9 | **Rata-rata Ping per Kapal** | Rata-rata jumlah baris per kapal | total baris / jumlah MMSI | Mengukur kepadatan data per kapal |
| 10 | **Schema & Range Pass Rate** | Persentase baris yang lolos batas fisik (lat ±90, lon ±180, sog ≤ 70) | jumlah baris valid / total × 100 | Menjamin data dalam rentang wajar |
| 11 | **Cakupan Sel H3 (res7)** | Jumlah sel H3 resolusi 7 yang tercakup (≈ 1 km²) | count distinct(H3\_int\_index\_7) | Menilai luas jangkauan spasial data |
| 12 | **Pelabuhan Tercover per Hari** | Rata-rata jumlah pelabuhan berbeda yang terdeteksi setiap hari | avg(count distinct(port\_id) per hari) | Mengukur variasi dan jangkauan port-call harian |
| 13 | **Inside-Port Share** | Proporsi titik AIS yang berada dalam poligon pelabuhan | (jumlah titik dalam pelabuhan / total titik) × 100 | Mengidentifikasi aktivitas pelabuhan yang relevan |
| 14 | **Koordinat Tidak Masuk Akal** | Draught = 0 dan sog = 0 (biasanya noise di darat) | (jumlah noise seperti ini / total) × 100 | Deteksi noise GPS atau spoofing |
| 15 | **Teleport (Jump > 30km/min)** | Persentase titik dengan loncatan spasial tak masuk akal | jarak / waktu > 30 km/min | Menyaring anomali spasial yang merusak trajektori |
| 16 | **Gap Sinyal > 60 menit** | Persentase jeda waktu antar ping terlalu lama | % ∆t antar ping > 60 menit | Mendeteksi blank spot atau AIS nonaktif |
| 17 | **Gap SOG vs Perhitungan** | Rata-rata selisih kecepatan SOG vs hasil perhitungan |  | sog - (jarak / ∆t) |
| 18 | **Konflik Metadata Kapal** | Jumlah MMSI dengan metadata ganda (IMO, nama, ukuran) | nunique(field) > 1 per MMSI | Menandakan kesalahan merge atau spoofing data |
| 19 | **Ingest Latency p95** | Latensi 95% data masuk (detik) dari waktu posisi ke waktu sistem | p95(dt\_insert - dt\_pos) | Menilai kedekatan data dengan real-time |
| 20 | **Duplikat Primary Key** | Persentase baris dengan duplikat (mmsi, dt\_pos\_utc) | jumlah duplikat PK / total × 100 | Menandakan replay data atau error ingest |
| 21 | **NavStatus ≠ SOG** | Persentase baris yang statusnya 'Moored/Anchored' tapi SOG-nya tinggi | (jumlah tidak konsisten / total) × 100 | Validasi konsistensi antara status & kecepatan |