PENJELASAN LENGKAP MENGENAI PENGGUNAAN DATA, FEATURE ENGINEERING, DAN FEATURE SELECTION

1. Mengapa Menggunakan Sample Data (Padahal Tidak Menggunakan Sample)?

Script tidak melakukan sampling apa pun. Seluruh data yang valid digunakan 100%.

Dataset 1 (Pharmacy)

- Total digunakan: 21.176 sampel (100% data valid)

Dataset 2 (Wave)

- Total digunakan: 7.870 sampel (100% data valid)

Proses Data Cleaning (Bukan Sampling)

Data yang dibuang berasal dari proses peningkatan kualitas:

- Missing values: data kosong atau corrupted
- Outliers (top 1%): anomali ekstrem yang dapat memicu bias
- Zero variance columns: kolom tanpa informasi
- Rows dengan lag = NaN: terjadi pada hari-hari awal karena belum ada riwayat

Contoh pada Dataset 1:

- Raw data: 24.748 baris
- Setelah cleaning: 21.176 baris
- 3.572 baris awal dibuang karena tidak memiliki fitur lag

---

2. Objektif dan Harapan Feature Selection

DATASET 1: PHARMACY TRANSACTION

Objektif Bisnis

Memprediksi demand obat untuk optimasi inventory management.

Target Variable

- `demand_class` (binary classification)
  - 0 = Low demand (≤ median)
  - 1 = High demand (> median)

Feature Engineering (46 fitur)

Termasuk:

- Temporal: day, month, quarter, day_of_week, is_weekend
- Lag features: qty_lag_1, qty_lag_7, qty_lag_14, ..., qty_lag_28
- Rolling statistics: qty_roll_mean_x, qty_roll_std_x, qty_roll_max_x
- EWMA: qty_ewma_3, qty_ewma_7, qty_ewma_14
- Change metrics: qty_change_1, qty_change_3, qty_change_7
- Coefficient of Variation: qty_cv_7, qty_cv_14, qty_cv_30

Tujuan Feature Selection

Mengidentifikasi fitur yang paling berpengaruh untuk prediksi demand.

Hasil Evaluasi Feature Selection

Metode	Fitur Terpilih	F1-Score	Perubahan
Baseline	46	0.8887	-
RFECV	6	0.9139	+2.83%
Information Gain	10	0.9052	+1.85%
Pearson	10	0.8532	-3.99%
Chi-Square	10	0.6515	-26.69%

Best Features (RFECV – 6 fitur)

No	Feature Name	Deskripsi	Fungsi Bisnis	Interpretasi
1	`qty_lag_1`	Quantity 1 hari yang lalu	Melihat pola demand kemarin sebagai baseline prediksi hari ini	Jika kemarin tinggi, kemungkinan besar hari ini juga tinggi (momentum)
2	`qty_roll_min_3`	Minimum quantity dalam 3 hari terakhir	Mendeteksi level demand terendah dalam periode pendek	Menandakan "lantai" demand - berguna untuk safety stock
3	`qty_ewma_3`	Exponential Weighted Moving Average 3 hari	Trend demand terkini dengan bobot lebih besar pada data terbaru	Lebih responsif terhadap perubahan demand dibanding simple average
4	`qty_change_1`	Persentase perubahan demand dari 1 hari lalu	Mengukur momentum pertumbuhan/penurunan demand	Jika +20%, berarti demand naik 20% dari kemarin (trend naik)

No	Feature Name	Deskripsi	Fungsi Bisnis	Interpretasi
5	`qty_change_3`	Persentase perubahan demand dari 3 hari lalu	Mengukur trend perubahan dalam jangka pendek	Mendeteksi pola naik/turun dalam rentang 3 hari
6	`qty_cv_7`	Coefficient of Variation demand 7 hari	Mengukur volatilitas/stabilitas demand mingguan	CV tinggi = demand tidak stabil, perlu buffer stock lebih besar

📊 **Penjelasan Tambahan:**

- **Lag features** (`qty_lag_1`): Nilai historis murni tanpa transformasi
- **Rolling statistics** (`qty_roll_min_3`): Agregasi dalam window waktu tertentu
- **EWMA** (`qty_ewma_3`): Moving average yang memberikan bobot eksponensial pada data terbaru
- **Change** (`qty_change_1`, `qty_change_3`): Rate of change untuk menangkap momentum
- **CV** (`qty_cv_7`): Standar deviasi / mean, mengukur variabilitas relatif

💡 **Insight Kunci:**

- Semua 6 fitur fokus pada **short-term patterns** (1-7 hari)
- Kombinasi **level** (lag, min), **trend** (change), dan **volatility** (cv) memberikan gambaran lengkap
- **TIDAK ADA** fitur temporal (day, month) yang terpilih → Pattern demand lebih dipengaruhi oleh riwayat bukan musim

Insight Bisnis

- Riwayat jangka pendek (1–3 hari) jauh lebih penting dibanding jangka panjang.
- Variabilitas dan perubahan tren lebih informatif daripada nilai absolut.
- Reduksi fitur dari 46 menjadi 6 meningkatkan performa model dan mempermudah deployment.

DATASET 2: WAVE MEASUREMENT

Objektif Bisnis

Memprediksi kondisi gelombang tinggi untuk keselamatan pelayaran dan operasi maritim.

Target Variable

- `wave_class` (binary classification)
  - 0 = Low wave (≤ 0.27 m)
  - 1 = High wave (> 0.27 m)

Features (13 oceanographic parameters)

No	Feature Name	Satuan	Deskripsi	Fungsi dalam Prediksi Gelombang	Range Tipikal

No	Feature Name	Satuan	Deskripsi	Fungsi dalam Prediksi Gelombang	Range Tipikal
1	`Hmax(m)`	meter	Maximum wave height - Tinggi gelombang maksimum	**Indikator utama** tinggi gelombang, berkorelasi langsung dengan target	0.1 - 5.0 m
2	`WaveDir(deg)`	derajat	Wave direction - Arah datangnya gelombang	Menentukan apakah gelombang datang dari laut lepas (berbahaya) atau pantai	0° - 360°
3	`WavePeriod(s)`	detik	Wave period - Jarak waktu antar puncak gelombang	Period panjang = gelombang besar energi tinggi = berbahaya	2 - 20 s
4	`WindSpeed(knots)`	knot	Wind speed - Kecepatan angin	Angin kencang membangkitkan gelombang tinggi	0 - 40 knots
5	`WindDir(deg)`	derajat	Wind direction - Arah angin	Arah angin mempengaruhi pembentukan dan arah gelombang	0° - 360°
6	`PrimSwell(m)`	meter	Primary swell height - Tinggi gelombang swell utama	Swell adalah gelombang dari badai jauh, bisa tiba-tiba tinggi	0.1 - 3.0 m
7	`WindSeaDir(deg)`	derajat	Wind sea direction - Arah gelombang yang dibangkitkan angin lokal	Berbeda dengan wave direction, ini khusus gelombang lokal	0° - 360°
8	`SurfCurrentDir`	derajat	Surface current direction - Arah arus permukaan	Arus kuat dapat memperbesar tinggi gelombang efektif	0° - 360°
9	`SeaSurfaceSalinity(PSU)`	PSU	Sea surface salinity - Kadar garam air laut	Mempengaruhi densitas air dan karakteristik gelombang	30 - 35 PSU
10	`SeaSurfaceTemp(°C)`	°C	Sea surface temperature - Suhu permukaan laut	Perbedaan suhu mempengaruhi tekanan udara dan pembentukan angin	25 - 32°C

No	Feature Name	Satuan	Deskripsi	Fungsi dalam Prediksi Gelombang	Range Tipikal
11	`AirPressure(hPa)`	hPa	Atmospheric pressure - Tekanan udara	Tekanan rendah = cuaca buruk = gelombang tinggi	990 - 1020 hPa
12	`Humidity(%)`	persen	Air humidity - Kelembaban udara	Indikator kondisi cuaca dan potensi hujan/badai	60 - 95%
13	`Visibility(km)`	km	Visibility - Jarak pandang	Visibilitas rendah = cuaca buruk = gelombang berpotensi tinggi	1 - 50 km

Tujuan Feature Selection

Menentukan parameter oseanografi yang paling berpengaruh terhadap prediksi gelombang.

Hasil Evaluasi

Metode	Fitur Terpilih	F1-Score	Perubahan
Baseline	13	0.9686	-
RFECV	5	0.9729	+0.44%
Pearson	7	0.9692	+0.07%
Information Gain	7	0.9682	-0.05%
Chi-Square	7	0.9650	-0.37%

Best Features (RFECV – 5 fitur)

No	Feature Name	Importance Score	Fungsi Kritis	Alasan Terpilih	Dampak Operasional
1	`Hmax(m)`	0.862 ★	**KUNCI UTAMA** - Tinggi gelombang maksimum	Berkorelasi langsung dengan target (0.86), dipilih semua metode	Threshold utama: Hmax > 0.27m = BAHAYA
2	`WaveDir(deg)`	0.735	Arah datang gelombang	Gelombang dari laut lepas lebih berbahaya dari pantai	Kapal harus hindari gelombang frontal
3	`WavePeriod(s)`	0.645	Periode/frekuensi gelombang	Period panjang = energi tinggi = gelombang destruktif	Period > 10s = gelombang berbahaya

No	Feature Name	Importance Score	Fungsi Kritis	Alasan Terpilih	Dampak Operasional
4	`SurfCurrentDir`	1.000	Arah arus permukaan	Arus berlawanan arah kapal memperbesar efek gelombang	Rute kapal harus sesuai arah arus
5	`WindSpeed(knots)`	0.687	Kecepatan angin	Angin kencang membangkitkan gelombang tinggi	Wind > 20 knots = peringatan cuaca

📊 **Analisis Feature Importance:**

- `Hmax(m)` mendominasi dengan score 0.862 → **MUST MONITOR**
- 4 dari 5 fitur terkait **karakteristik gelombang langsung** (Hmax, WaveDir, WavePeriod, SurfCurrent)
- Hanya 1 fitur atmosfer (`WindSpeed`) yang terpilih → **Kondisi laut > kondisi udara**

💡 **Implikasi Praktis:**

1. **Sensor Prioritas**: Pasang sensor Hmax, WaveDir, WavePeriod, Current, WindSpeed
2. **Early Warning System**: Monitor Hmax realtime, jika > 0.27m → alert otomatis
3. **Routing Decision**: Kombinasi WaveDir + SurfCurrentDir tentukan jalur aman kapal
4. **Cost Saving**: Tidak perlu monitor 13 sensor, cukup 5 sensor kritis

🚨 **Decision Rules Sederhana:**

```
IF Hmax > 0.27m AND WavePeriod > 10s AND WindSpeed > 20 knots:
    → CANCEL SAILING (High Risk)
ELIF Hmax > 0.27m OR (WaveDir = Frontal AND SurfCurrentDir = Opposite):
    → DELAY SAILING (Medium Risk)
ELSE:
    → SAFE TO SAIL (Low Risk)
```

Insight Bisnis

- Hmax adalah fitur paling kritis (dipilih oleh seluruh metode).
- Arah gelombang lebih berpengaruh dibanding kecepatan angin.
- Monitoring dapat difokuskan pada 5 sensor sehingga lebih efisien.

KESIMPULAN UMUM

Manfaat Feature Selection

Efisiensi Komputasi

- Dataset 1: 46 → 6 fitur (reduksi 87%)

- Dataset 2: 13 → 5 fitur (reduksi 62%)
- Training dan inference lebih cepat

Peningkatan Performa

- Dataset 1: peningkatan F1 +2.83%
- Dataset 2: peningkatan F1 +0.44%

Interpretability

- Model lebih mudah dijelaskan
- Memfokuskan analisis pada fitur yang benar-benar relevan

Cost Reduction

- Dataset 2: cukup monitor 5 sensor
- Dataset 1: cukup melacak 6 metrik utama

Mengurangi Overfitting

- Fitur lebih sedikit → generalisasi lebih baik

Takeaway

Feature selection bukan tentang mengurangi data, tetapi menemukan fitur yang paling relevan untuk membangun model yang lebih efisien, akurat, dan mudah dipahami.