

# Proyecto 1: Estación de llenado y taponado

Python para ingenieros introductorio

Máster Universitario en Informática Industrial y Robótica

Pedro Antonio Toledo Delgado

2025

## Resumen

Este documento define un proyecto práctico orientado a consolidar competencias de programación en Python (fundamentos, NumPy, Pandas y Matplotlib) mediante el análisis de una línea sencilla de *llenado y taponado* de botellas. El sistema cuenta con telemetría a 1 Hz, eventos discretos y registros por ciclo (botella). El objetivo es construir un *pipeline reproducible* de ingesta de datos, cálculo de indicadores (KPIs, OEE), análisis numérico (correlaciones y regresión OLS en NumPy) y visualización técnica, concluyendo con un informe sintético de hallazgos y recomendaciones.

## 1. Descripción operativa del sistema

La línea realiza el **llenado de botellas** mediante una bomba dosificadora y el **taponado** posterior. Las botellas se desplazan sobre una *cinta transportadora* a velocidad ajustable. Se trabajan dos formatos habituales (250 ml y 500 ml) y se registran *micro-paradas* debidas a atascos u operaciones de rutina (p. ej., limpieza o cambio de formato).

### Flujo de proceso

1. **Alimentación y transporte:** las botellas entran a la célula y avanzan sobre la cinta.
2. **Llenado:** una bomba dosificadora inyecta el producto a caudal controlado, durante el tiempo necesario para alcanzar la masa/volumen objetivo.
3. **Taponado:** un sistema aprieta-tapones aplica un par controlado para garantizar cierre y estanqueidad.
4. **Salida e inspección:** se verifica el peso llenado y se clasifica cada unidad como OK o NG según tolerancias.

### Actuadores principales

- **Bomba dosificadora (variador de frecuencia).** Regula el *caudal* de llenado. El caudal efectivo depende de la consigna y de la **temperatura del producto** (a mayor temperatura, menor viscosidad y mayor caudal real). Es habitual aplicar rampas suaves y tiempos de estabilización para evitar sobre/ subllenado.
- **Cinta transportadora.** Ajusta la *velocidad de línea*; determina el *tiempo de residencia* bajo el punto de llenado y condiciona el *tiempo de ciclo* por botella. En *RUN* se sitúa típicamente entre 0,22 m/s y 0,38 m/s según formato; en *STOP* la velocidad es cero.

- **Apreta-tapones.** Controla el *par de apriete* para asegurar la calidad del cierre sin dañar el tapón ni la rosca. El par objetivo puede diferir por formato y material de tapón. Un par insuficiente puede provocar NG por fuga; un exceso puede generar defectos mecánicos.

### Sensores y señales (telemetría 1 Hz)

- **temp\_prod** ( $^{\circ}\text{C}$ ). Temperatura del producto a la entrada de la dosificación. Afecta directamente a la *viscosidad* y, por tanto, al caudal real. Suele presentar deriva lenta y pequeñas oscilaciones por control térmico.
- **vel\_cinta** ( $\text{m s}^{-1}$ ). Velocidad instantánea de la cinta (0 en parada). Sirve como señal primaria para discriminar estados *RUN/STOP* y se correlaciona con el *tiempo de ciclo*.
- **caudal** ( $\text{ml s}^{-1}$ ). Caudal instantáneo medido o estimado. En *STOP* debe ser cero; en marcha varía con consigna y temperatura. Puede filtrarse o presentar cuantización por el método de medida.
- **energia\_kwh** (kWh, acumulada). Lectura acumulativa del consumo eléctrico de la célula. Permite derivar *potencia* por diferencias; la señal debe ser *no decreciente* salvo pequeños dientes de sierra por resolución del contador.

### Datos por botella (evento/ciclo)

- **ts\_ciclo** (ISO 8601). Marca temporal del fin de ciclo de una botella (punto inmediatamente posterior al llenado o tras la verificación de peso).
- **id\_botella**. Identificador único correlativo o de trazabilidad.
- **formato\_ml**  $\in \{250, 500\}$ . Tamaño de la botella; condiciona consignas de caudal y ritmos de línea.
- **tiempo\_ciclo\_s**. Tiempo transcurrido entre dos **ts\_ciclo** consecutivos. Si no está presente, se obtiene por diferencia temporal ordenando por **ts\_ciclo**.
- **peso\_lleno\_g**. Masa de producto dispensada. Se compara con el objetivo del formato (densidad  $\approx 1 \text{ g/ml}$ ) para evaluar tolerancia.
- **ok\_ng**  $\in \{\text{OK}, \text{NG}\}$ . Resultado de inspección; típico umbral de aceptación:  $\pm 2\%$  respecto al peso objetivo, pudiendo incluir reglas adicionales (fugas, tapón).

### Eventos operativos

- **micro\_parada**. Interrupción breve (decenas de segundos a pocos minutos) por atasco o ajuste menor. Durante la ventana del evento, la **vel\_cinta** es cero y el **caudal** se inhibe; afecta a la disponibilidad (componente *A* del OEE).
- **cambio\_formato**. Operación programada para alternar entre 250 ml y 500 ml. Implica parada controlada para reajustes mecánicos y de consignas; tras el evento cambian los parámetros nominales de velocidad y dosificación.
- **limpieza**. Ciclo de *CIP* (limpieza en sitio) o mantenimiento de higiene. Supone tiempo de parada prolongado y puede modificar transitoriamente señales térmicas y de consumo.

## 2. Datos y estructura de ficheros

El conjunto de datos se entrega en tres ficheros CSV ubicados en `data/`. Todos los ficheros están codificados en UTF-8, usan coma como separador, punto como separador decimal y una fila inicial de cabeceras. Las marcas temporales se expresan en formato ISO 8601 con zona UTC (sufijo Z). La ventana temporal de análisis viene dada por las primeras y últimas marcas temporales presentes en `telemetry.csv`.

### `telemetry.csv` (1–2 días, 1 Hz)

Este fichero contiene la telemetría continua de la línea a 1 Hz. Cada fila corresponde a una muestra temporal y describe el estado operativo en ese instante.

- **ts** (*timestamp ISO 8601, UTC*). Marca temporal monótona no decreciente.
- **temp\_prod** ( $^{\circ}\text{C}$ ). Temperatura del producto a la entrada de dosificación; afecta a la viscosidad y, por tanto, al caudal efectivo. Suele mostrar deriva lenta y pequeñas oscilaciones de control. Rango orientativo: 18–35  $^{\circ}\text{C}$ .
- **vel\_cinta** ( $\text{m s}^{-1}$ ). Velocidad instantánea de la cinta. Toma el valor 0 en parada y típicamente 0.22–0.38 m/s en marcha, dependiendo del formato.
- **caudal** ( $\text{ml s}^{-1}$ ). Caudal instantáneo de dosificación. Debe ser 0 en parada; en marcha varía con la consigna y con **temp\_prod**. Puede presentar cuantización o filtrado según el método de medida. Rango orientativo: 0–12 ml/s.
- **energia\_kwh** (kWh, acumulada). Energía eléctrica acumulada desde el inicio de la campaña. Debe ser no decreciente salvo pequeños dientes de sierra atribuibles a resolución del contador. Permite derivar potencia por diferencias.

**Integridad esperada:** marcas temporales ordenadas, sin duplicados exactos; **energia\_kwh** prácticamente monótona; **vel\_cinta**=0 implica **caudal**=0. Pueden existir huecos breves de muestreo (segundos), que deberán tratarse en la fase de ingestión.

### `eventos.csv`

Relación de eventos discretos con inicio y fin, utilizados para contextualizar la telemetría y distinguir periodos de *RUN/STOP*, cambios de formato y tareas de limpieza.

- **ts\_ini, ts\_fin** (*timestamp ISO 8601, UTC*). Límites temporales del evento; se asume **ts\_ini** < **ts\_fin**. Para el mapeo con telemetría se recomienda el intervalo semiabierto [**ts\_ini**, **ts\_fin**).
- **tipo**  $\in \{\text{micro\_parada}, \text{cambio\_formato}, \text{limpieza}\}$ . Clasificación operativa del evento.
- **detalle**. Texto breve con información adicional (p. ej., causa de la micro-parada o destino del cambio).

Los intervalos de **micro\_parada** anulan la producción y deben reflejarse como *STOP* en el estado operativo. Los **cambio\_formato** implican parada controlada y modificación posterior de consignas. Las **limpieza** suponen paradas más prolongadas y pueden afectar transitoriamente a temperatura y consumo.

## botellas.csv

Registros por ciclo (una fila por botella) que recogen la trazabilidad unitaria, el resultado de calidad y, cuando procede, el tiempo de ciclo asociado.

- **ts\_ciclo** (*timestamp ISO 8601, UTC*). Marca temporal del fin de ciclo de una unidad (tras llenado y/o verificación).
- **id\_botella**. Identificador único (correlativo o de trazabilidad).
- **formato\_ml**  $\in \{250, 500\}$ . Formato de la unidad; condiciona consignas de caudal y ritmo de línea.
- **tiempo\_ciclo\_s** (s). Duración del ciclo. Si no estuviera presente, puede obtenerse por diferencias sucesivas de **ts\_ciclo** tras ordenar por tiempo.
- **peso\_lleno\_g** (g). Masa de producto dispensada. Se compara con el objetivo del formato (densidad  $\approx 1$  g/ml) para decidir conformidad.
- **ok\_ng**  $\in \{OK, NG\}$ . Resultado de inspección; la regla de aceptación típica es  $\pm 2\%$  respecto al objetivo, pudiendo coexistir reglas adicionales (fugas, cierre).

**Consistencia esperada:** unicidad de **id\_botella**; no decreciente en **ts\_ciclo**; valores de **formato\_ml** coherentes con los intervalos de **cambio\_formato** de **eventos.csv**.

## Coherencia e integración entre ficheros

El análisis integra las tres fuentes sobre el eje temporal:

1. **Telemetría**  $\leftrightarrow$  **Eventos**: cada muestra de **telemetria.csv** se etiqueta con el estado operativo (*RUN/STOP*) en función de si su **ts** cae dentro de algún intervalo de **micro\_parada** o de paradas por **cambio\_formato/limpieza**. Esto permite calcular disponibilidad y tiempos efectivos de marcha.
2. **Telemetría**  $\leftrightarrow$  **Botellas**: los registros de **botellas.csv** se asocian a la telemetría por proximidad temporal (misma ventana o mediante *join* por minuto) para estudiar relaciones entre **temp\_prod**, **caudal** y los resultados de calidad (**ok\_ng**, **peso\_lleno\_g**) y para estimar el **tiempo\_ciclo\_s** cuando sea necesario.
3. **Agregación temporal**: la telemetría se resume a resolución de 1 min (medias, percentiles, %STOP) para su comparación con KPIs por hora/turno y con métricas unitarias agregadas (p.ej., *Wh/ud*).

## Diccionario de datos resumido

Fichero	Columna	Unidad/tipo	Descripción
telemetria.csv	ts	ISO 8601 (UTC)	Marca temporal (1 Hz).
	temp_prod	°C (float)	Temperatura del producto; 18–35 °C.
	vel_cinta	ms <sup>-1</sup> (float)	Velocidad de cinta; 0 en parada.
	caudal	mls <sup>-1</sup> (float)	Caudal instantáneo; 0 en parada.
	energia_kwh	kWh (float, acumulada)	Energía acumulada (no decreciente).
eventos.csv	ts_ini, ts_fin	ISO 8601 (UTC)	Intervalo del evento ([ini, fin)).
	tipo	categórica	micro_parada, cambio_formato, limpieza.
	detalle	texto	Información adicional del evento.
botellas.csv	ts_ciclo	ISO 8601 (UTC)	Fin de ciclo por unidad.
	id_botella	entero/str	Identificador único.
	formato_ml	{250, 500}	Formato de botella.
	tiempo_ciclo_s	s (float)	Duración del ciclo (derivable por diferencias).
	peso_lleno_g	g (float)	Masa dispensada; comparación con objetivo.
	ok_ng	OK/NG	Resultado de inspección de calidad.

### 3. Plan de trabajo y requisitos funcionales

El proyecto se organiza en fases encadenadas y orientadas a reproducibilidad. Se emplea Python con NumPy, Pandas y Matplotlib; todo el código deberá ser determinista cuando aplique (semillas fijadas), con funciones puras y artefactos intermedios versionables.

#### Fase 1 — Ingesta y validación (Pandas)

**Objetivo** Construir una capa de datos fiable que garantice tipados, orden temporal, coherencia física de señales y una etiqueta operativa *RUN/STOP* utilizable en fases posteriores.

**Entradas** telemetria.csv, eventos.csv, botellas.csv en la carpeta data/, con marcas temporales ISO 8601 en UTC (Z).

#### Salidas

- df\_tel, df\_evt, df\_pz tipados y ordenados, con índice temporal en UTC.
- Columna estado en df\_tel (RUN/STOP) y, opcionalmente, %STOP por minuto.
- Informe breve de calidad de datos (conteos, huecos, atípicos, correcciones).

#### Procedimiento

1. **Carga y tipado:** leer CSV con dtype explícitos; convertir columnas ts\* a datetime64[ns, UTC]; establecer ts como índice en telemetría.
2. **Orden y duplicados:** ordenar por tiempo; eliminar duplicados exactos conservando el primero; validar monotonía no estricta del índice.
3. **Validaciones de rango** (marcar, no eliminar):
  - temp\_prod ∈ [18, 35] °C; vel\_cinta ∈ [0, 0,5] m/s; caudal ∈ [0, 12] ml/s.
  - energia\_kwh no decreciente salvo cuantización; ver paso 4.
4. **Monotonicidad de energía:** calcular  $\Delta E_i = E_i - E_{i-1}$ . Para  $\Delta E_i < 0$ , fijar a 0 (*clip*) y reconstruir  $E$  acumulado; registrar el número de correcciones.
5. **Frecuencia y huecos:** confirmar frecuencia nominal de 1 Hz. Reindexar a rejilla de 1 s y:

- para huecos  $\leq 10$  s: interpolación lineal en **temp\_prod** y **caudal**; **vel\_cinta** por *forward-fill* con límite de ventana;
  - para huecos  $> 10$  s: marcar segmento inválido (bandera booleana) para su exclusión en KPIs sensibles.
6. **Atípicos** (marcado): z-score  $|z| > 3$  o IQR ( $Q11.5 \cdot IQR$ ,  $Q3+1.5 \cdot IQR$ ) en **temp\_prod** y **caudal**. Incluir columna **es\_atipico** sin suprimir observaciones.
7. **Etiqueta RUN/STOP por segundo**: construir una máscara de parada a partir de **eventos.csv** (**micro\_parada**, **cambio\_formato**, **limpieza**) usando el intervalo  $[ts\_ini, ts\_fin]$ . Definir:

$$RUN\_vel_i := (vel\_cinta_i \geq 0,05 \text{ m/s}), \quad STOP\_evt_i := ts_i \in \text{intervalos de evento},$$

$$estado_i := \begin{cases} RUN & \text{si } RUN\_vel_i \wedge \neg STOP\_evt_i, \\ STOP & \text{en otro caso.} \end{cases}$$

(Opcional) aplicar histéresis de 2–3 s al cambio de estado para evitar oscilaciones.

8. **Agregación a 1 minuto** (para diagnóstico temprano): por minuto calcular **temp\_mean**, **temp\_p95**, **caudal\_mean**, **%STOP** (segundos en **STOP/60**). Conservar también los segundos **RUN** para disponibilidad.

### Criterios de aceptación

- Sin duplicados temporales; índices monótonos; tipados coherentes.
- **energia\_kwh** no decreciente tras corrección; registro del número de *clips*.
- **vel\_cinta=0**  $\Rightarrow$  **estado=STOP** y **caudal=0** en esos instantes.
- Huecos  $\leq 10$  s interpolados y etiquetados; huecos mayores marcados para exclusión.
- Informe de calidad generado con recuentos de huecos, atípicos y correcciones.

### Fase 2 — Ingeniería de variables y KPIs (NumPy + Pandas)

**Propósito** Derivar variables físicas a partir de la telemetría por segundo y calcular indicadores clave de desempeño (KPIs) por ventanas temporales (minuto, hora, turno) de forma vectorizada y reproducible.

#### Definiciones y notación

- $t_i$ : marca temporal del instante  $i$  (en segundos, ordenado).  $\Delta t_i = \frac{t_i - t_{i-1}}{3600}$  (horas entre  $i-1$  e  $i$ ).
- $E_i$  (kWh): energía acumulada medida en  $t_i$  (no decreciente salvo cuantización).
- $P_i$  (kW): potencia media en el intervalo  $(t_{i-1}, t_i]$ .
- $estado_i \in \{RUN, STOP\}$ : etiqueta operativa del segundo  $i$  (según Fase 1).
- $W$ : ventana de agregación (p. ej., una hora o un turno).
- $N_W = OK_W + NG_W$ : número total de unidades en  $W$  (suma de conforme/no conforme).
- $\Delta E_{kWh}(W)$ : incremento de energía acumulada dentro de  $W$  (kWh).

- $t_{\text{plan}}(W)$ : tiempo planificado de producción en  $W$  (horas).
- $t_{\text{RUN}}(W)$ : tiempo efectivo en marcha en  $W$  (horas), obtenido sumando segundos con **RUN**.
- $f \in \{250, 500\}$ : formato de botella (ml). Masa objetivo  $m_{\text{obj}}(f) \approx f$  (g), asumiendo densidad  $\approx 1$  g/ml.
- $t_{\text{nom}}(f)$ : tiempo de ciclo nominal por formato (p.ej., 1,8 s para 250 ml, 2,2 s para 500 ml).
- $t_{\text{medio\_RUN}}(W)$ : tiempo de ciclo medio observado en **RUN** dentro de  $W$  (segundos).

**Potencia instantánea a partir de energía acumulada** Para cada intervalo  $(t_{i-1}, t_i]$  y energía acumulada  $E_i$ :

$$\Delta E_i = \max\{E_i - E_{i-1}, 0\}, \quad \Delta t_i = \frac{t_i - t_{i-1}}{3600} \text{ (h)}, \quad P_{\text{kW},i} = \frac{\Delta E_i}{\Delta t_i}, \quad P_{\text{W},i} = 1000 P_{\text{kW},i}.$$

(Se recomienda suavizado opcional por media móvil para mitigar cuantización).

**Agregación a 1 minuto (telemetría)** Para cada minuto  $m$ :

$$\text{temp\_mean}(m) = \text{mean}(T), \quad \text{temp\_p95}(m) = \text{p95}(T), \quad \text{caudal\_mean}(m) = \text{mean}(q), \quad \text{P\_kW\_mean}(m) = \text{mean}(P_{\text{kW}})$$

$$\% \text{STOP}(m) = 100 \cdot \frac{\#\{i \in m : \text{estado}_i = \text{STOP}\}}{60}.$$

Estas series minuto servirán como base para KPIs horarios/por turno.

**Objetivo de masa por formato**

$$m_{\text{obj}}(250) = 250 \text{ g}, \quad m_{\text{obj}}(500) = 500 \text{ g}.$$

Con tolerancia típica del  $\pm 2\%$ , una unidad está dentro de tolerancia si

$$|\text{peso\_lleno\_g} - m_{\text{obj}}(f)| \leq 0,02 \cdot m_{\text{obj}}(f).$$

**KPIs por hora y por turno** Sea  $W$  la ventana (hora/turno). Definiciones consistentes:

- **Throughput** (ud/h):

$$\text{Throughput}(W) = \frac{N_W}{\text{horas}(W)}.$$

- **Scrap** (% no conforme):

$$\text{Scrap}(W) = 100 \cdot \frac{NG_W}{OK_W + NG_W} \quad (\text{si } N_W > 0; \text{ en otro caso NaN}).$$

- **Tiempo en marcha** (h):

$$\text{Tiempo en marcha}(W) = t_{\text{RUN}}(W).$$

- **Energía específica** (Wh/ud):

$$\text{Wh/ud}(W) = \frac{1000 \cdot \Delta E_{\text{kWh}}(W)}{N_W} \quad (\text{si } N_W > 0; \text{ en otro caso NaN}).$$

- **% dentro de tolerancia:**

$$\% \text{Tol}(W) = 100 \cdot \frac{\#\{\text{unidades en tolerancia}\}}{N_W} \quad (\text{si } N_W > 0).$$

**OEE (Overall Equipment Effectiveness)** Usar una única definición coherente en todo el análisis. Se proponen dos formulaciones equivalentes; se recomienda elegir *una* y mantenerla:

**(A) Definición por tiempos y ciclo nominal**

$$\text{Availability}(W) = \frac{t_{\text{RUN}}(W)}{t_{\text{plan}}(W)}, \quad \text{Performance}(W) \approx \frac{\overline{t_{\text{nom}}}(W)}{t_{\text{medio\_RUN}}(W)}, \quad \text{Quality}(W) = \frac{OK_W}{OK_W + NG_W},$$

$$\text{OEE}(W) = \text{Availability} \cdot \text{Performance} \cdot \text{Quality}.$$

Aquí  $\overline{t_{\text{nom}}}(W)$  es el tiempo de ciclo nominal *ponderado por formato* en  $W$ .

**(B) Definición por ritmos real y teórico**

$$\text{Performance}(W) = \frac{\frac{N_W}{t_{\text{plan}}(W)}}{\frac{N_{\text{teo}}(W)}{t_{\text{plan}}(W)}}, \quad N_{\text{teo}}(W) = \sum_f \frac{\text{tiempo planificado en } f}{t_{\text{nom}}(f)}.$$

Con esta formulación, Availability y Quality son idénticas a las anteriores.

**Consideraciones de implementación**

- Tratar *divisiones por cero*: si  $N_W = 0$  o  $t_{\text{RUN}}(W) = 0$ , devolver NaN y anotar el caso.
- En ventanas con múltiples formatos, ponderar  $t_{\text{nom}}(f)$  por el número de unidades o por el tiempo planificado por formato.
- Preferir operaciones vectorizadas (sin bucles Python): `groupby-agg`, `resample`, y máscaras booleanas.

**Fase 3 — Análisis numérico (NumPy puro)**

**Propósito** Cuantificar relaciones entre variables operativas, modelar el *error de llenado* y detectar situaciones anómalas mediante reglas simples, usando álgebra lineal y operaciones vectorizadas en NumPy (sin librerías de ML).

**Preparación y notación**

- Sea  $t$  el tiempo; trabajar en una **rejilla temporal consistente**: minuto (agregando telemetría) o por-ciclo (alineando cada botella con la muestra de telemetría más próxima). Mantener esta decisión en toda la fase.
- Variables continuas: temperatura  $T$  (`temp_prod`), caudal  $q$  (`caudal`), potencia  $P$  (`P_kW`), tiempo de ciclo  $t_c$  (`tiempo_ciclo_s`).
- Variable binaria:  $\text{RUN} \in \{0, 1\}$  (1 si en marcha).
- **Masa objetivo**  $m_{\text{obj}}(f)$ : 250 g o 500 g según formato  $f$ .
- **Error de llenado**:  $e = \text{peso\_lleno\_g} - m_{\text{obj}}(f)$ .
- Al calcular métricas, usar máscaras para ignorar NaN y mantener el mismo subconjunto de índices en todas las variables implicadas.

**Correlaciones de Pearson** Sea  $\mathbf{x}, \mathbf{y}$  un par de series (mismo tamaño tras enmascarar NaN). La **correlación de Pearson** se define como

$$r_{xy} = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}.$$

Calcular la matriz  $r$  para  $\{T, q, P, t_c, e\}$  sobre la rejilla elegida. *Buenas prácticas*: estandarizar (restar media y dividir por desviación típica) antes de componer la matriz, y reportar también el número de muestras válidas por par.

**Regresión lineal OLS en NumPy (modelo para  $e$ )** **Objetivo**: explicar el error de llenado  $e$  con predictores físicos.

$$e = \beta_0 + \beta_1 (T - 25) + \beta_2 q + \beta_3 \text{RUN} + \varepsilon.$$

**Diseño**:

$$\mathbf{X} = \begin{bmatrix} 1 & T_1 - 25 & q_1 & \text{RUN}_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_n - 25 & q_n & \text{RUN}_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}.$$

**Estimación** (dos opciones equivalentes en NumPy):

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad \text{o} \quad \hat{\beta} = \arg \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2 \quad (\text{np.linalg.lstsq, más estable numéricamente}).$$

**Calidad de ajuste**:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}, \quad R^2 = 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2}, \quad R_{\text{aj}}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1},$$

donde  $p$  es el número de predictores (sin contar el intercepto). **Diagnósticos mínimos**: media de residuos  $\approx 0$ , varianza de residuos razonable, ausencia de tendencia en residuos vs. predicción. **Signos esperados**:  $\beta_1 < 0$  (a mayor  $T$  baja la viscosidad y disminuye el sesgo de subllenado),  $\beta_2 < 0$  (más caudal reduce error negativo),  $\beta_3$  cercano a 0 si ya se filtra por RUN, aunque puede capturar cambios de régimen.

**Detección de anomalías (reglas operativas)**

1. **Subllenado sostenido** (por-ciclo o minuto). Definir indicador  $s_i = \mathbf{1}\{e_i < -0.02 \cdot m_{\text{obj}}(f_i)\}$ . Contabilizar segundos consecutivos en  $s$  mediante *convolución* con un núcleo de unos:

$$c_i = (s * \mathbf{1}_L)_i, \quad L = 60 \text{ (seg)}.$$

Señalar intervalos donde  $c_i \geq L$  como anomalías de subllenado sostenido; combinar índices contiguos en ventanas  $[\text{ini}, \text{fin}]$ .

2. **Operación ineficiente** (por hora). Para cada hora  $h$ , calcular  $\text{Wh/ud}(h) = \frac{1000 \cdot \Delta E_{\text{kWh}}(h)}{N_h}$  (si  $N_h > 0$ ). Definir umbral  $\tau = \text{p95}\{\text{Wh/ud}(h)\}$  y marcar horas con  $\text{Wh/ud}(h) > \tau$ . (Alternativa robusta: usar MAD alrededor de la mediana).

**Salidas esperadas**

- Matriz de correlaciones (Pearson) y tabla con el número de muestras válidas por par.
- Vector  $\hat{\beta}$ ,  $R^2$  y  $R_{\text{aj}}^2$  del modelo para  $e$ , junto con un breve comentario sobre el signo y magnitud de cada coeficiente.
- Listado de intervalos anómalos: *subllenado sostenido* y *operación ineficiente*, con **ts\_ini**, **ts\_fin** y motivo.

## Fase 4 — Visualización (Matplotlib)

**Objetivo** Comunicar de forma clara el comportamiento de la línea y sus efectos en los KPIs. Las figuras deben ser reproducibles, legibles y con unidades en ejes y leyendas.

### Estándares de presentación

- Guardado doble formato: PNG (150–200 DPI) y SVG vectorial en `fig/`.
- Tamaño sugerido:  $10 \times 4$  in (series temporales) y  $8 \times 4$  in (barras/histogramas).
- Etiquetas completas (variable y unidad), cuadrícula discreta, ejes bien acotados.
- Tiempo formateado con `DateFormatter` y `HourLocator/MinuteLocator`; `tight_layout()` para evitar solapes.
- Leyenda única por figura; si es necesario, ubicarla fuera del área de trazado (`bbox_to_anchor`).

### Figuras requeridas

1. **Serie temporal 12–24 h (twinx)**. Eje izquierdo: `temp_prod` (línea). Añadir banda  $\pm\sigma$  con desviación estándar móvil (ventana 10 min) usando `fill_between`. Eje derecho (twinx): `caudal` (línea). Sombrear intervalos STOP (de `eventos.csv`) con `axvspan`. Opcional: líneas verticales finas en `cambio_formato/limpieza`. *Claves*: eje X compartido, unidades en ambos ejes Y, leyenda combinada, y formato horario legible.
2. **Barras apiladas: OEE por turno (A/P/Q)**. Para cada turno (M, T, N) representar barras apiladas con *Availability*, *Performance* y *Quality* en %. Mostrar etiquetas % en cada segmento (centradas) y el valor de OEE total encima de la barra. *Claves*: eje Y en 0–100 %, orden consistente (A abajo  $\rightarrow$  Q arriba), anotación opcional con  $N$  de unidades y  $t_{\text{RUN}}$ .
3. **Histograma del error de llenado  $e$  por formato**. Dos subgráficos (o superposición con transparencia): 250 ml y 500 ml. Selección de *bins* con la regla de Freedman–Diaconis ( $h = 2 \text{IQR } n^{-1/3}$ ) para comparar colas. Marcar verticalmente media y p95; sombrear banda de tolerancia  $\pm 2\%$  alrededor de  $m_{\text{obj}}(f)$ . *Claves*: indicar  $n$  por formato, eje X en gramos, comentar asimetrías (sub/sobre-llenado).
4. **Scatter “binned”: temperatura vs. % en tolerancia**. Agrupar por `temp_prod` en bins de  $0.5^\circ\text{C}$ . Para cada bin, calcular % de unidades dentro de tolerancia y `caudal_mean`. Representar el punto en el centro del bin (*scatter*) con color según `caudal_mean` (añadir `colorbar`); tamaño opcional proporcional al conteo del bin. *Claves*: mostrar intervalo de confianza binomial (p. ej., Wilson) como barra vertical; limitar a bins con  $n$  suficiente (p. ej.,  $n \geq 30$ ).
5. **Wh/ud por hora con hitos operativos**. Barras por hora con  $\text{Wh/ud} = \frac{1000 \Delta E_{\text{kWh}}}{N}$ . Añadir línea horizontal en p95 como umbral de *ineficiencia*. Trazar líneas verticales en `cambio_formato` y anotar su hora. Opcional: superponer una línea secundaria con *Throughput* (eje derecho) para leer consumo específico vs. carga. *Claves*: ocultar horas sin producción ( $N = 0$ ) o marcarlas explícitamente; ordenar etiquetas de tiempo de forma compacta.

### Buenas prácticas específicas

- **Escalas y límites**: evitar autoscaling extremo; fijar márgenes razonables para destacar la banda  $\pm\sigma$  y las líneas de referencia.
- **Unidades siempre visibles**: incluir unidad en los ejes ( $^\circ\text{C}$ , ml/s, kW, Wh/ud, %).

- **Anotaciones sobrias:** usar `annotate` sólo para hitos clave (p95, media, cambios de formato).
- **Consistencia visual:** misma tipografía/tamaño; mismas convenciones de símbolos para todas las figuras del informe.

## Fase 5 — Informe breve

**Objetivo** Sintetizar en 1–2 páginas la evidencia cuantitativa obtenida, destacando implicaciones operativas y el impacto esperado sobre OEE y consumo específico (Wh/ud).

### Estructura sugerida (concisa y accionable)

1. **Resumen ejecutivo** (5–7 líneas). Qué se analizó, ventana temporal y *dos* mensajes clave.
2. **Contexto y alcance.** Fuentes (**telemetria, eventos, botellas**), supuestos y exclusiones.
3. **KPIs principales** (tabla breve). OEE por turno, Scrap (%), Throughput (ud/h), Wh/ud (p95, mediana).
4. **Hallazgos** (3–5 puntos, con cifras). Ej.: “Por debajo de 22 °C la probabilidad de subllenado crece  $\times 1,8$ ”.
5. **Recomendaciones** (2–3, *SMART* y cuantificadas). Ej.: “Fijar consigna a  $26 \pm 0,5$  °C  $\Rightarrow$  -0.9 pp Scrap; habilitar standby cuando  $N = 0$  durante  $>15$  min  $\Rightarrow$  -8 % Wh/ud”.
6. **Impacto estimado.** Cambio esperado en *Availability/Performance/Quality*, OEE total y Wh/ud, con breve método de estimación.
7. **Riesgos y límites.** Calidad de datos, sensibilidad a parámetros (bins, umbrales), supuestos de linealidad del modelo.

### Figuras recomendadas (2–3)

- Serie temporal (temp/caudal con STOP sombreado).
- Barras apiladas OEE por turno (A/P/Q).
- Histograma del error de llenado  $e$  por formato con banda de tolerancia  $\pm 2\%$ .

Incluir pies de figura autoexplicativos y unidades. Todas las imágenes deben referenciarse desde `fig/`.

## 4. Fórmulas clave

**Notación**  $t_i$ : marca temporal (s).  $E_i$ : energía acumulada en  $t_i$  (kWh).  $P_i$ : potencia media en  $(t_{i-1}, t_i]$  (kW).  $N = OK + NG$ .  $t_{\text{plan}}$ : tiempo planificado (h).  $t_{\text{RUN}}$ : tiempo en marcha (h).  $t_{\text{nom}}$ : tiempo de ciclo nominal (s).  $t_{\text{medio\_RUN}}$ : ciclo medio observado en RUN (s).  $\Delta E_{\text{kWh}}(W)$ : energía consumida en ventana  $W$  (kWh).  $y$ : variable objetivo (p.ej.,  $e$ ).  $\mathbf{X}$ : matriz de diseño.  $\beta$ : coeficientes.

### Balances y potencias

$$\Delta E_{\text{kWh},i} = \max(E_i - E_{i-1}, 0), \quad \Delta t_{h,i} = \frac{t_i - t_{i-1}}{3600}, \quad P_{\text{kW},i} = \frac{\Delta E_{\text{kWh},i}}{\Delta t_{h,i}}.$$

## KPIs de consumo y producción

$$\text{Wh/ud}(W) = \frac{1000 \cdot \Delta E_{\text{kWh}}(W)}{N_W}, \quad (N_W > 0)$$
$$\text{Throughput}(W) = \frac{N_W}{\text{horas}(W)}.$$

## Componentes de OEE

$$\text{Availability}(W) = \frac{t_{\text{RUN}}(W)}{t_{\text{plan}}(W)},$$
$$\text{Performance}(W) \approx \frac{\overline{t_{\text{nom}}}(W)}{t_{\text{medio\_RUN}}(W)},$$
$$\text{Quality}(W) = \frac{OK_W}{OK_W + NG_W},$$
$$\text{OEE}(W) = \text{Availability} \cdot \text{Performance} \cdot \text{Quality}.$$

Alternativa de *Performance* por ritmos:

$$\text{Performance}(W) = \frac{\frac{N_W}{t_{\text{plan}}(W)}}{\frac{N_{\text{teo}}(W)}{t_{\text{plan}}(W)}}, \quad N_{\text{teo}}(W) = \sum_f \frac{\text{tiempo planificado en } f}{t_{\text{nom}}(f)}.$$

## Regresión lineal (OLS) y métricas

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (\text{equivalente numéricamente a } \text{np.linalg.lstsq}),$$
$$R^2 = 1 - \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2}, \quad R_{\text{aj}}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1}.$$

## Reglas de calidad y tolerancia

$$\text{Conforme} \iff |\text{peso\_lleno\_g} - m_{\text{obj}}(f)| \leq 0,02 \cdot m_{\text{obj}}(f).$$

# 5. Entregables esperados

## Estructura mínima del repositorio

```
.
data/                # CSV entregados (solo lectura)
fig/                 # Gráficos generados (PNG + SVG)
src/                 # Funciones reutilizables (opcional)
proyecto_llenado.ipynb
features_1min.parquet # Agregación a 1 min (si aplica)
informe.md | informe.ipynb
README.md            # Instrucciones de ejecución (paso a paso)
requirements.txt     # Versionado de dependencias
```

## Listado de ficheros

1. `proyecto_llenado.ipynb` (o estructura `src/` + un notebook de reporte).

2. `features_1min.parquet`, tablas por hora/turno y figuras exportadas en `fig/` (PNG a 150–200 DPI y SVG).
3. `informe.md` o `informe.ipynb` (exportable a PDF) con la estructura de la Fase 5.
4. `README.md` con: versiones de Python/paquetes, pasos de ejecución, orden de celdas y notas de reproducibilidad (semillas, parámetros).

### **Criterios de aceptación**

- Reproducibilidad: ejecutar el notebook genera los mismos artefactos (mismas semillas y rutas relativas).
- Claridad: figuras legibles con unidades; tablas con totales y % explícitos.
- Cohesión: KPIs, figuras e informe refieren a la misma ventana temporal y a las mismas definiciones.