# LABORATORIO 3: Preprocesamiento, EDA, Estacionalidad y Predicción en Series Temporales (DEIS, SINCA, energía, etc.)

## T – TASK (Definición de Tarea Aplicada)

**Contexto.** Cada equipo elegirá una **serie temporal real** (Chile u otra fuente pública) y construirá un flujo reproducible para: **(i)** preparar datos temporales, **(ii)** hacer EDA específico, **(iii)** analizar estacionalidad/tendencia y **(iv) pronosticar** con validación temporal.

**Ejemplos de fuentes (no excluyentes):**

- **Salud:** ingresos UCI / hospitalizaciones (DEIS).
- **Ambiente:** $PM_{2.5}$ / $PM_{10}$, $O_3$, $NO_2$ (SINCA).
- **Energía:** demanda/consumo eléctrico, generación.
- **Otros:** transporte, comercio, clima, etc.

**Objetivos del laboratorio (alineados a RA1, RA2, RA3, RA5):**

- **RA1 – Calidad/estructura temporal.** Auditar fechas, frecuencia, duplicados, huecos, outliers; justificar que la serie sirve para el problema.
- **RA2 – ETL + EDA específico.** Pipeline reproducible (carga→limpieza→transformación "tidy"), EDA con foco temporal y descomposición.
- **RA3 – Modelado y predicción.** Baselines + 1–2 modelos; análisis de estacionalidad; validación **respetando el tiempo**; métricas por horizonte e **intervalos**.
- **RA5 – Reproducibilidad.** Entrega ejecutable con README y bitácora de decisiones.

**Alcance y restricciones:**

- Serie univariada principal para pronóstico; se permiten exógenas (feriados, clima, precio) si se justifican.
- Mínimos orientativos: mensual ≥ 36 obs; semanal ≥ 104; diaria ≥ 365.
- **Prohibido** el *random split*; usar **holdout temporal**.

## I – INFORMATION (Recursos y Metodología)

**Datos.** Ustedes descargan y citan su fuente (DEIS, SINCA, coordinadores/ministerios, etc.). Documenten fecha de descarga y diccionarios.

**Herramientas sugeridas.** Python (pandas, numpy, matplotlib), statsmodels (ARIMA/SARIMAX, ETS), pmdarima/Prophet (opcional), ydata-profiling (opcional), git.

**Marco metodológico sugerido.**

1. **Ingesta y auditoría temporal (DQR-T).**
    - Parseo de fechas, *timezone* si aplica, definir **frecuencia** (D/W/M), ordenar.
    - Duplicados, **huecos**, valores atípicos, cambios de método.
    - Re-muestreo/agrupación (p.ej., de horario→diario, diario→semanal) con justificación.
2. **EDA específico de series.**
    - Gráficos de **nivel** y de **cambios** (diferencias o retornos).
    - **Descomposición** (p.ej., STL): tendencia, estacionalidad, residuo.
    - Plots estacionales (por mes / día-semana), feriados/eventos.
3. **Estacionariedad y preparación.**
    - Si la amplitud crece: **log/Box-Cox**.
    - **Diferencia mínima necesaria** ($\Delta$ y $\Delta_s$) para estabilizar; mostrar **antes/después**.

4. **Modelado y predicción:**
    - **Baselines obligatorios:** Naïve y **Seasonal Naïve**.
    - 1–2 modelos candidatos: **SARIMA/SARIMAX**, **ETS/estado-espacio**, **TBATS** (múltiples estacionalidades) o **Prophet** (opcional).
    - Si usan exógenas: justificar causalidad/calendario y evitar *leakage*.
5. **Validación temporal (opcional)**
    - **Holdout final** + **rolling/expanding origin**.
    - Reportar **MAE y RMSE por horizonte** (ej.: h=1, h=4, h=12 según frecuencia) e **intervalos de predicción**.
    - Comparar contra baselines; si no los superan, discutir por qué.
6. **Documentación.**
    - Scripts/notebooks modulares, **README** reproducible, *requirements* o *environment*.

## L – LEARNING (Proceso de Aprendizaje Estructurado)

**Fase 1 — Selección y DQR-T (RA1) – 20%**
• Elección justificada del dataset. • Perfil temporal (frecuencia, huecos, outliers, riesgos/sesgos).

**Fase 2 — ETL + EDA con descomposición (RA2) – 25%**
• Pipeline ejecutable y *tidy*. • Gráficos claros (nivel/cambios, estacionalidad).

**Fase 3 — Estacionariedad y preparación (RA2) – 10%**
• Transformaciones (log/Δ/Δ_s) motivadas y evidenciadas con “antes/después”.

**Fase 4 — Modelado y Predicción (RA3) – 30%**
• Baselines + 1–2 modelos. • Hiper-parámetros documentados. • **Intervalos** y lectura de resultados.

**Fase 5 — Reproducibilidad (RA5) – 15%**
• Repo ejecutable, README, bitácora de decisiones.

# T – TRANSFER (Entregables y Formato)

**1) Informe breve (4–6 págs.)**

- **Contexto y fuente** (qué mide, periodo, frecuencia, por qué es relevante).
- **DQR-T** (calidad temporal: huecos, outliers, cambios de método).
- **EDA** (nivel/cambios, descomposición, estacionalidad).
- **Preparación** (transformaciones y justificación).
- **Modelos y validación** (baselines, candidatos, métricas por horizonte, intervalos).
- **Conclusiones** (qué aprendieron, limitaciones, próximos pasos).

**2) Repositorio reproducible (zip o enlace).**
Estructura sugerida:

data/ (raw/ processed/)
src/ (etl.py, features.py, model.py, eval.py)
notebooks/
reports/ (figures/, tables/)
README.md requirements.txt|environment.yml LICENSE

**3) "One-pager" (1 slide) con lo mejor del laboratorio.**

- **3 gráficos** (p.ej., descomposición, plot estacional, pronóstico con intervalos) + **3 insights**.

**Criterios de aceptación mínimos.**

- Pipeline **end-to-end** que corre y genera dataset "tidy".
- **Validación temporal correcta** (sin *random split*), **baselines incluidos**.
- Gráficos legibles (ejes/leyendas), interpretación concisa.
- Cita de la **fuente de datos** y fecha de descarga.

**Entrega y plazos:** *12/09/2025 a las 23:59*.

# RÚBRICA DE EVALUACIÓN (Lab 3)

**Excelencia Técnica – RA3 (30%)**

- Baselines bien implementados y superados o discutidos; modelos adecuados; intervalos de predicción y lectura crítica.

**EDA y Preparación – RA2 (35%)**

- ETL correcto, descomposición y análisis de estacionalidad claros; transformaciones justificadas.

**Rigor de Calidad de Datos – RA1 (20%)**

- DQR temporal completo (frecuencia, huecos, outliers, cambios de método) y riesgos/sesgos.

**Reproducibilidad y Documentación – RA5 (15%)**

- Repo ejecutable, README claro, bitácora de decisiones, orden profesional.

## Nota para los estudiantes

**Reglas de oro:**

1. **Ordena el tiempo** (frecuencia definida, sin huecos “fantasma”).
2. **Dibuja antes de modelar** (nivel y cambios).
3. **Valida mirando hacia adelante** (holdout + rolling).
4. **Siempre** compite contra un **naïve**.
5. Entrega **intervalos** y explica **supuestos**.