

LABORATORIO REDISEÑADO: ANALÍTICA APLICADA EN SALUD Y EPIDEMIOLOGÍA

T – TASK (Definición de Tarea Aplicada)

Contexto Ejecutivo:

Desarrollar una solución analítica **reproducible** que transforme datos públicos de salud en **insights accionables** para apoyar decisiones en salud pública/medicina de precisión. El foco es conectar **EDA + inferencia estadística + modelado** con una **comunicación clara** para audiencias clínicas/gestión.

Objetivos Específicos:

- Construir un pipeline con auditoría de calidad, tratamiento de faltantes y outliers.
- Realizar EDA riguroso y pruebas/inferencia (intervalos de confianza, contraste de hipótesis).
- Entrenar y diagnosticar modelos estadísticos (p. ej., regresión/GLM) y/o series de tiempo (ARIMA/forecast) cuando aplique.
- Entregar visualizaciones y reportes que traduzcan resultados técnicos en recomendaciones para política/gestión clínica.
- Documentar decisiones éticas y limitaciones (sesgos, validez externa, trazabilidad).

Líneas/Temas de Proyecto (elegir uno o proponer variante fundada):

1. **Cáncer de origen desconocido (CUP)** en Chile (GRD/FONASA 2019–2024).
2. **Complicaciones post-trasplante renal pediátrico** (<15 años) y estancias/egresos (FONASA).
3. **Cáncer materno y resultados al nacimiento** (CIE-10 C* y O00–O9A; peso RN, condición de alta).
4. **Contaminación ambiental y enfermedades respiratorias** (SINCA + FONASA/DEIS).

5. **Contaminación ambiental y enfermedades crónicas no transmisibles** (SINCA + FONASA/DEIS).

I – INFORMATION (Recursos y Metodología)

Datasets Base (obligatorios, se pueden enriquecer):

- **GRD/FONASA 2019–2024** (egresos, diagnósticos/procedimientos).
<https://public.tableau.com/views/PropuestaTableroGRD/PropuestaTableroGRD?%3AshowVizHome=no#1>
- **SINCA (MMA)**: contaminantes ambientales/monitoreo.
<https://sinca.mma.gob.cl/index.php/>
- **DEIS / Superintendencia de Salud**: egresos/indicadores públicos.

Herramientas Técnicas:

- **Python** (Pandas, NumPy, SciPy), **visualización** (Matplotlib/Seaborn),
- **statsmodels** (regresión/GLM, pruebas), **scikit-learn** (métricas/modelos básicos),
- **ARIMA/Prophet** (si hay series), control de versión (**git**), opcional **MLflow/Docker**.

Marco Metodológico:

1. **EDA + DQR**: perfilado, faltantes, atípicos (IQR/MAD), diccionario de variables.
2. **Inferencia**: hipótesis, IC, pruebas adecuadas al diseño y a la distribución.
3. **Modelado**: especificación, **diagnóstico de supuestos** (linealidad, colinealidad, homocedasticidad, residuos), sensibilidad.
4. **Forecast (opcional)**: estacionariedad, ACF/PACF, validación temporal.
5. **Comunicación**: narrativa técnica + ejecutiva; decisiones éticas y de trazabilidad.

L – LEARNING (Proceso de Aprendizaje Estructurado)

Fase 1 — Auditoría & EDA (20%)

- DQR: integridad, consistencia, reglas de negocio.
- EDA visual/numérica; definición de variables y cohortes.

Fase 2 — Inferencia & Especificación del Modelo (25%)

- Preguntas/hipótesis, pruebas estadísticas, IC.
- Diseño del modelo (regresión/GLM) y/o protocolo de series.

Fase 3 — Implementación & Diagnóstico (35%)

- Pipeline ETL reproducible; tratamiento de outliers/faltantes.
- Entrenamiento, métricas (RMSE/MAE/AIC/BIC, o AUROC/F1 si aplica),
- Diagnóstico de supuestos y análisis de sensibilidad.

Fase 4 — Comunicación, Ética & Transferencia (20%)

- Visualizaciones claras; reporte ejecutivo con recomendaciones.
- Riesgos/sesgos, limitaciones, validez externa y plan de mejora.

T – TRANSFER (Transferencia y Aplicación)

Entregables (con formato y evaluación):

- **Reporte Ejecutivo (30%)** — 6–8 páginas para decisores (hallazgos, implicancias, recomendaciones priorizadas).
- **Informe Técnico (30%)** — 10–15 páginas (EDA, inferencia, modelos, diagnóstico, sensibilidad, límites).
- **Repositorio Reproducible (25%)** — Estructura `data/src/notebooks/reports`, README, instrucciones de corrida, versión de paquetes.
- **Póster A1 o 10 slides (15%)** — Visual storytelling del problema, método y hallazgos.

Criterios de Alineación a Resultados de Aprendizaje (RA):

- **RA1** Calidad/estructura de datos (DQR)
- **RA2** Pipeline ETL reproducible (limpieza, outliers, imputación, features)
- **RA3** Modelos e inferencia con diagnóstico de supuestos
- **RA4** Series de tiempo/forecast cuando corresponda
- **RA5** Documentación/versionado (reproducibilidad)
- **RA6** Comunicación ética y efectiva (técnica/ejecutiva)

RÚBRICA DE EVALUACIÓN ACADÉMICA

Excelencia Técnica (40%)

- Pipeline reproducible; tratamiento correcto de datos.
- Modelos adecuados al problema; diagnóstico de supuestos.
- Métricas correctas y análisis de sensibilidad.

Rigor Analítico (30%)

- Inferencia bien planteada (hipótesis/IC/pruebas).

- Coherencia entre EDA→modelo→conclusiones; límites explícitos.
- Uso apropiado de validación (temporal si aplica).

Aplicabilidad y Transferencia (20%)

- Claridad del valor práctico (políticas/gestión clínica).
- Recomendaciones priorizadas y trazables a la evidencia.
- Repositorio ejecutable por terceros.

Ética y Profesionalismo (10%)

- Identificación de sesgos, privacidad y uso responsable.
- Propuestas de mitigación y documentación de decisiones.
- Orden, prolijidad y cumplimiento de formato.

Entrega y Fechas:

- **Propuesta (E0):** semana 3
- **Avance E1–E2:** semana 5
- **Avance E3:** semana 6
- **Entrega final (todos los entregables):** semana 7

Nota: Se permite incorporar fuentes adicionales siempre que se documenten su calidad, sesgos y compatibilidad. La reproducibilidad es requisito mínimo de aprobación.