Tecnicatura Superior en Ciencia de Datos e Inteligencia Artificial

Centro Politécnico Superior Malvinas Argentinas

ML Auditoría de Prestaciones Extrahospitalarias (TDF)

Entrega 2 – Descripción del Dataset y Origen

Materia: Aprendizaje Automático

Docente: Nicolás Caballero

Estudiante: Nancy Julieta Cassano

Fecha: 20/10/2025

1. Resumen

Se presenta un dataset sintético diseñado para entrenar modelos supervisados capaces de predecir la decisión "Autorizar / No autorizar" una prestación extrahospitalaria solicitada para pacientes sin cobertura médica en la provincia de Tierra del Fuego.

El conjunto de datos replica las reglas administrativas y clínicas reales aplicadas por la Dirección de Prestaciones Médicas, garantizando privacidad, reproducibilidad y cumplimiento normativo.

2. Objetivo del dataset

Proveer una base de datos estructurada, verosímil y completamente anónima para:

- Entrenar y comparar clasificadores (Regresión Logística, Árboles, Random Forest).
- Evaluar el efecto del balanceo de clases sobre precisión y recall.
- Ajustar umbrales de decisión que minimicen falsos negativos en casos clínicos críticos.
- Facilitar la documentación y validación del proceso de auditoría automatizada.

3. Origen y método de construcción

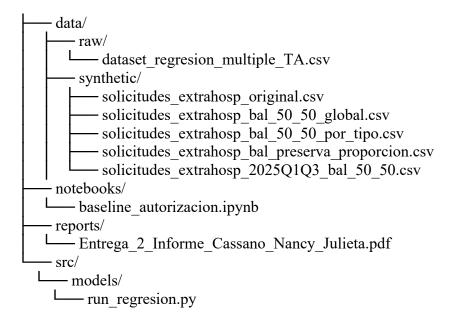
- Fuente: datos generados sintéticamente, sin información personal real.
- Reglas de negocio simuladas:
 - Domicilio en Tierra del Fuego.
 - Ausencia de obra social o prepaga activa.
 - Verificación de derechos potenciales (monotributo, cónyuge, etc.).
 - Evaluación socioeconómica (ingresos, red de apoyo, informe social).
 - Disponibilidad de la prestación en el sistema público provincial.
 - Documentación obligatoria: orden médica, DNI, estudios complementarios.

- Variable objetivo (autorizar): 1 si cumple criterios completos / 0 en caso contrario.
- Motivo de generación sintética: resguardar identidad de pacientes, controlar desbalance y garantizar replicabilidad para fines académicos.

4. Licencias

- Código: MIT License.
- **Datos:** Creative Commons CC0 1.0 Dominio Público. Ambas licencias permiten libre uso con fines educativos y de investigación.

5. Estructura del repositorio



6. Versión y volumen del dataset

- **Instancias:** 1200 solicitudes simuladas.
- Variables: 22 columnas sociodemográficas, clínicas y administrativas.
- Ventana temporal (variante): 2025-01-01 a 2025-09-30.
- Clase objetivo: desbalance real (autorizar > no autorizar).

7. Variantes del dataset

- 1. Original (desbalanceado) refleja prevalencias reales.
- 2. Balanceado 50/50 global igual número de clases.
- 3. Balanceado 50/50 por tipo de prestación equilibrio intra-categorías.
- 4. Balanceado preservando proporciones reales mixto.
- 5. Acotado 2025-Q1–Q3 (50/50) control temporal para validación cruzada.

8. Diccionario de variables

Variable	Tipo	Valores	Descripción
id_solicitud	String	SOL-xxxxx	Identificador único simulado
fecha_solicitud	Date	AAAA-MM-DD	Fecha de ingreso
zona	Categórica	sur / norte	Jurisdicción operativa
hospital_origen	Categórica	HRU / HCN / CAPS_*	Establecimiento solicitante
edad	Numérica	0–100	Edad del paciente
ingresos_mensuales	Numérica	Pesos	Ingresos declarados
cobertura	Binaria	0 / 1	Obra social activa
derecho_cobertura	Binaria	0 / 1	Derecho potencial a cobertura
disp_publica	Binaria	0 / 1	Prestación disponible en red pública
tipo_prestacion	Categórica	prácticas / estudios / insumos	Tipo de prestación
documentacion_completa	Binaria	0 / 1	Documentación completa
informe_social	Categórica	completo / incompleto	Calidad del informe social
residencia_tdf	Binaria	0 / 1	Domicilio en TDF verificado
prioridad	Categórica	urgente / programable	Urgencia clínica
autorizar	Label	0 / 1	Variable objetivo

9. Calidad de datos y validaciones

Durante la generación del dataset se aplicaron controles básicos de coherencia y calidad, con el fin de asegurar que los valores sean consistentes y representen adecuadamente las condiciones reales del proceso de autorización.

Entre las principales validaciones realizadas se incluyen:

- Verificación de rangos: se controlaron valores atípicos o fuera de rango en variables como edad e ingresos mensuales.
- Coherencia interna: se revisó la consistencia entre cobertura y derecho_cobertura, evitando combinaciones lógicamente imposibles.
- Relación entre variables: se observa que las solicitudes con documentación incompleta presentan una menor probabilidad de ser autorizadas.
- Duplicados y valores faltantes: se eliminaron registros repetidos y se revisó la ausencia de valores nulos en campos clave.

- Reproducibilidad: se fijaron semillas aleatorias para mantener la trazabilidad de los resultados en futuras ejecuciones.

10. Limitaciones y sesgos

Si bien el dataset fue diseñado para reflejar situaciones reales de solicitud y autorización de prestaciones, presenta algunas limitaciones propias de su carácter sintético:

- Naturaleza simulada: al no provenir de casos reales, no contempla situaciones excepcionales o atípicas, como intervenciones judiciales o traslados de urgencia.
- Simplificación de reglas: las condiciones aplicadas pueden representar criterios administrativos más rígidos que los utilizados en la práctica cotidiana.
- Recorte temporal: algunas versiones del dataset se acotan al período enero—septiembre de 2025, lo que podría introducir un sesgo temporal en los análisis.
- Alcance delimitado: no se incluyen variables económicas ni de facturación, dado que estos aspectos quedan fuera del objetivo de esta etapa del proyecto.