

Descripción del Dataset Procesado

Resumen General

Este archivo describe el **dataset final preprocesado** resultante del proyecto de análisis predictivo del Trastorno del Espectro Autista (TEA) en niños. Tras un exhaustivo proceso de limpieza, transformación y depuración, se obtuvo un conjunto de datos adaptado para el modelado con algoritmos de aprendizaje automático.

Cambios Realizados al Dataset Original

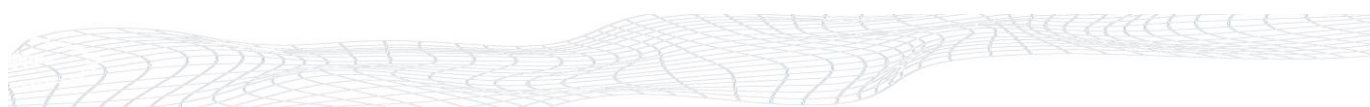
- **Renombrado de Columnas**

Se tradujeron los nombres de las variables para facilitar la interpretación.

Evaluación y Selección de Variables Relevantes

Durante la etapa de limpieza y análisis exploratorio del dataset, se evaluó la pertinencia de cada variable para el objetivo del proyecto.

- **Variables mantenidas:** fueron consideradas relevantes y útiles para el modelado predictivo las siguientes variables.
 - **A1_Score a A10_Score:** preguntas clave del test de cribado. Fundamentales para el modelo.
 - **Edad:** edad numérica del niño. Puede influir en la manifestación del TEA.
 - **Género:** género del niño. Diversos estudios lo relacionan con diferencias en prevalencia.
 - **Etnia:** grupo étnico. Puede aportar información sociodemográfica relevante.
 - **Ictericia_Neonatal:** historial de ictericia. Reportado como posible factor de riesgo.
 - **Antecedente_Familiar_TEA:** antecedentes familiares. Predictivamente valiosa.
 - **Clase_TEA:** variable objetivo.
- **Eliminación de Variables Irrelevantes o Problemáticas:** por razones de redundancia, alta cardinalidad o riesgo de sesgo, se eliminaron las siguientes variables.



- **Resultado_Diagnóstico**: cálculo derivado de las preguntas A1–A10. Altamente correlacionada con la variable objetivo; introducirla genera fuga de información (*data leakage*).
- **Grupo_Edad**: Redundante con Edad. Todos los casos pertenecen a un único grupo (4-11 years).
- **Relación_con_el_Niño**: No describe al niño, sino al encuestado; alto sesgo social.
- **País_de_Residencia**: Alta cardinalidad, poca representación por país. Inaplicable al contexto local (TDF).
- **Uso_App_Anterior**: Refiere al adulto encuestado, no aporta valor predictivo directo.
- **Etnia**: Falta de representatividad local, alta proporción de datos faltantes, posible sesgo.

Justificación completa incluida en el análisis exploratorio de la notebook.

• Tratamiento de Valores Nulos

- La variable Edad contenía 4 valores nulos.
- Se imputaron mediante la **mediana** de la columna (Edad = 6).

• Conversión y Codificación de Variables

- Edad: de float64 a int.
- Variables binarias (Género, Ictericia_Neonatal, Antecedente_Familiar_TEA, Clase_TEA) fueron convertidas a valores numéricos:
 - m/‘f’ → 0/1
 - no/yes → 0/1
 - NO/YES → 0/1

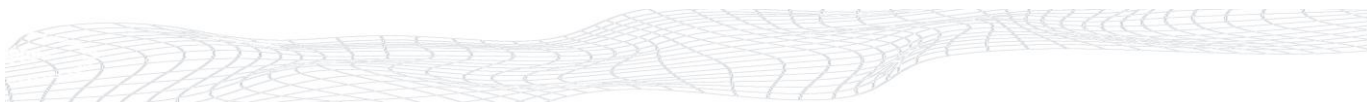
Descripción de las Variables Finales

Variable	Tipo	Descripción
A1_Score a A10_Score	Binaria	Preguntas conductuales en base al AQ-10 para niños
Edad	Entera	Edad del niño en años (entre 4 y 11)
Género	Binaria	0 = Masculino, 1 = Femenino
Ictericia_Neonatal	Binaria	Antecedentes de ictericia neonatal (0 = No, 1 = Sí)
Antecedente_Familiar_TEA	Binaria	Historia familiar de TEA (0 = No, 1 = Sí)
Clase_TEA	Binaria	Variable objetivo: 0 = No indicios de TEA, 1 = Posibles indicios de TEA

Referencia:

0 = masculino, 1 = femenino

0 = no, 1 = sí





**CENTRO POLITÉCNICO SUPERIOR
MALVINAS ARGENTINAS**

Formato Final del Dataset

- **Observaciones:** 292
 - **Variables:** 15 (todas numéricas)
 - **Sin valores nulos**
 - **Formato de guardado:** data/processed/
Dataset_Procesado_TEA_INFANCIA.csv
-

