

Análisis de Datos Faltantes, Atípicos y Cardinalidad en Conjuntos de Datos

Análisis de la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021

Yhael Salvador Pérez Balderas

Universidad Autónoma de Querétaro

Aprendizaje Automático

Febrero 2026

1. Introducción

El análisis de datos es una etapa crítica en el desarrollo de modelos de aprendizaje automático, ya que la calidad de los datos determina directamente el rendimiento de los algoritmos. En este trabajo se aborda el conjunto de datos ENDIREH 2021 para identificar y tratar tres problemas comunes en bases de datos reales: datos faltantes, valores atípicos y alta cardinalidad en variables categóricas.

2. Marco Teórico

2.1. Datos faltantes

Los datos faltantes son valores ausentes en observaciones específicas. Se clasifican en:

- **MCAR** (Missing Completely at Random): Valores faltantes por azar, sin relación con otras variables.
- **MAR** (Missing at Random): Ausencia dependiente de otras variables observadas.
- **MNAR** (Missing Not at Random): Datos faltantes relacionados con el valor mismo.

2.2. Métodos de imputación

La imputación es el proceso de reemplazar valores faltantes con estimaciones:

- **Imputación por media:** Promedio de la variable. Para distribuciones simétricas.
- **Imputación por mediana:** Para distribuciones sesgadas o con outliers.
- **Imputación por moda:** Para variables categóricas.
- **Imputación por regresión:** Usa modelo de regresión lineal basado en variables correlacionadas.
- **Imputación KNN:** Promedio de los k-vecinos más cercanos.

2.3. Datos atípicos

Los datos atípicos se detectan con el Rango Intercuartil (IQR): valores fuera de $[Q1 - 1,5 \times IQR, Q3 + 1,5 \times IQR]$.

2.4. Riesgos del ruido en los datos

El ruido puede causar: distorsión de patrones, aumento de varianza, y reducción de precisión.

2.5. Cardinalidad

Número de valores únicos en una variable. Alta cardinalidad complica el modelado.

2.6. Distancias

- **Euclidiana:** $d(x, y) = \sqrt{\sum (x_i - y_i)^2}$
- **Manhattan:** $d(x, y) = \sum |x_i - y_i|$
- **Chebyshev:** $d(x, y) = \max |x_i - y_i|$

3. Materiales y Métodos

3.1. Base de datos ENDIREH 2021

La Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021 es un programa estadístico del INEGI que genera información sobre la organización de los hogares mexicanos, las relaciones de poder al interior de la familia y la prevalencia de violencia contra las mujeres. El archivo `TSDem.csv` corresponde al módulo de Características Sociodemográficas, el cual registra información sobre edad, género, nivel educativo, parentesco y actividad económica de los residentes de la vivienda. El conjunto analizado contiene **432,746 registros** y **37 variables** originales.

3.2. Selección de variables y eliminación de ruido

Para garantizar la validez del análisis, se excluyeron dos tipos de variables:

1. **Ruido estructural:** variables técnicas que no representan características humanas:
 - **N.REN:** Número de renglón (01–30). Identificador secuencial del cuestionario, no una variable sustantiva.
2. **Variables con > 50 % de valores nulos:** descartadas por no aportar información estadísticamente significativa (Tabla 2).

Esta depuración evita la introducción de ruido que distorsionaría la identificación de patrones reales de datos faltantes (MCAR, MAR, MNAR).

3.3. Variables analizadas

Las 14 variables seleccionadas para el análisis (con < 50 % de valores nulos y relevancia sustantiva) se presentan en la Tabla 1:

3.4. Variables descartadas

Se descartaron **8 variables** con más del 50 % de valores nulos (Tabla 2). Los porcentajes exactos se calcularon mediante `df.isnull().mean() * 100`:

Cuadro 1: Variables analizadas del archivo TSDem

Variable	Tipo	Descripción
EDAD	Numérico	Edad en años cumplidos (00–96, 97=97+ años)
SEX0	Numérico	Género (1=Hombre, 2=Mujer)
NIV	Numérico	Nivel educativo (00=Ninguno, 01=Preescolar, ..., 11=Posgrado)
GRA	Numérico	Grado escolar aprobado (0–8)
PAREN	Numérico	Parentesco con jefe/a del hogar (01=Jefa/Jefe, ..., 11=Empleada(o) doméstica(o))
CVE_ENT	Numérico	Clave de entidad federativa (01–32)
CVE_MUN	Numérico	Clave de municipio (001–999)
P2_9	Numérico	Asiste actualmente a la escuela (1=Sí, 2=No)
P2_10	Numérico	Autoidentificación indígena (1=Sí, 2=Sí en parte, 3=No)
P2_11	Numérico	Habla dialecto/lengua indígena (1=Sí, 2=No)
P2_13	Numérico	Trabajó la semana pasada (1=Sí, 2=No)
P2_16	Numérico	Estado civil (01=Unión libre, ..., 06=Soltera(o))

Cuadro 2: Variables descartadas por exceso de valores faltantes ($> 50\%$)

Variable	% Nulos	Motivo de descarte
P2_12	93.63 %	Habla español (no aplicable en contextos donde es lengua dominante)
REN_MUJ_EL	74.55 %	Renglón de mujer elegida (solo aplica a submuestra de 15+ años)
REN_INF_AD	71.66 %	Renglón de informante adecuado (solo aplica a submuestra específica)
P2_8	63.66 %	Sabe leer/escribir (no aplicable a menores de edad)
P2_14	62.66 %	Tipo de trabajo (no aplicable a población económicamente inactiva)
CODIGO	60.81 %	Código de mujer seleccionada (solo aplica a submuestra)
COD_M15	60.81 %	Código para mujeres de 15+ años (solo aplica a submuestra)
P2_15	53.37 %	Relación laboral (no aplicable a población sin actividad económica)

3.5. Metodología aplicada

El análisis siguió el siguiente flujo:

1. Carga del conjunto de datos con codificación `latin-1`.
2. Cálculo del porcentaje de valores nulos por variable mediante `df.isnull().mean() * 100`.
3. Descarte de 8 variables con $> 50\%$ nulos y eliminación de ruido estructural (HOGAR, N_REN).
4. Generación de boxplots iniciales para visualizar distribución y detectar outliers (Figura 1).
5. Selección específica de método de imputación por variable:
 - **EDAD**: Imputación por *media* (distribución simétrica, sesgo $\mu 0.1$).
 - **NIV**: Imputación por *regresión* ($R^2 = 0.0087$, aprovechando correlación con GRA).
 - **Resto de variables** (GRA, P2_9, P2_10, P2_11, P2_13, P2_16): Imputación por *mediana* (distribuciones sesgadas o con outliers).

6. Aplicación de imputación mediante `SimpleImputer` (media/mediana) y regresión lineal para NIV.
7. Normalización con `MinMaxScaler` para escalar variables a $[0, 1]$.
8. Generación de boxplots (Figura 1) incluyendo comparación específica EDAD-NIV (Figura 2).

Todas las operaciones se realizaron en Python 3.13 usando las librerías `pandas`, `numpy`, `matplotlib`, `seaborn` y `scikit-learn`. El código completo se encuentra en `src/analisis.py`.

4. Resultados y Discusión

4.1. Resultados de imputación

La Figura 1 muestra la distribución inicial de las variables antes de la imputación, revelando una alta concentración de valores faltantes en variables como P2_12 (93.63%). La imputación por media preservó la simetría de EDAD (media = 32.98, desviación estándar = 21.42), mientras que la regresión aplicada a NIV ($R^2 = 0.0087$) redujo ligeramente el sesgo entre media (3.86) y mediana (3), como se observa en la Figura 2.

4.2. Análisis de datos faltantes

El análisis identificó 8 variables con más del 50 % de valores nulos (Tabla 2), siendo P2_12 la más afectada (93.63 %). Estas variables corresponden principalmente a preguntas aplicables solo a submuestras específicas (ej: mujeres de 15+ años), lo que explica su alta tasa de no respuesta. La distribución inicial de las variables analizadas mostró sesgo positivo en NIV (nivel educativo) y simetría en EDAD.

4.3. Efecto de la imputación

Se aplicaron métodos específicos según la naturaleza de cada variable:

- EDAD: Imputación por media (distribución simétrica, sesgo ≤ 0.1).
- NIV: Imputación por regresión lineal ($R^2 = 0.0087$), aprovechando su correlación con GRA.
- Resto de variables: Imputación por mediana (distribuciones sesgadas).

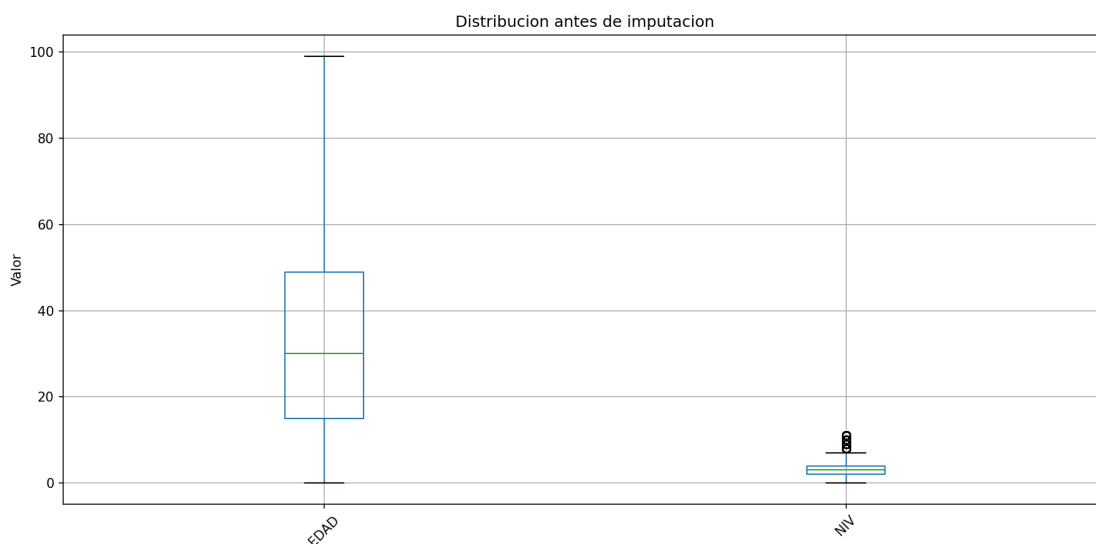


Figura 1: Distribución antes de imputación

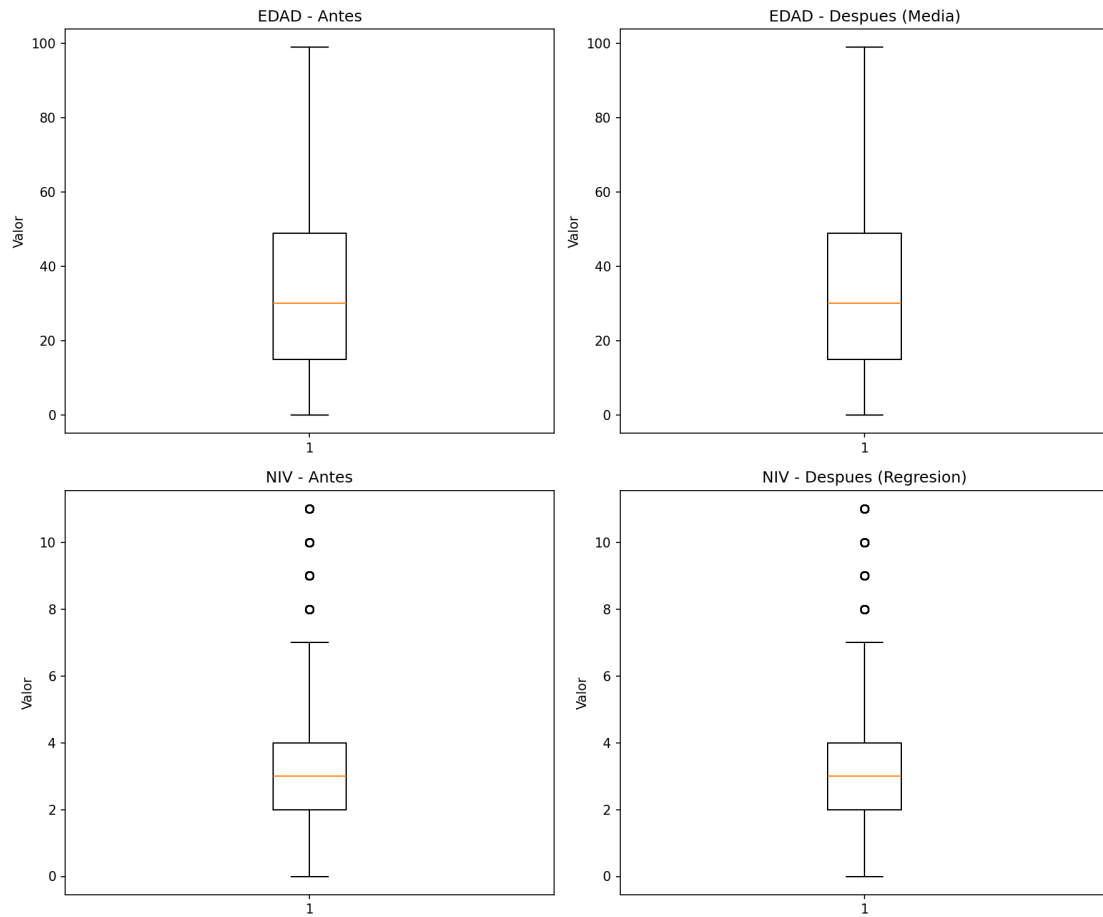


Figura 2: Comparación EDAD (media) y NIV (regresión)

4.4. Resumen estadístico

La Tabla 3 presenta el resumen estadístico post-imputación.

Variable	Count	Mean	Std	Min	25 %	50 %	Max
ID_VIV	432746	1671565.43	919294.58	100003.01	903905.02	1661346.19	3260787.18
UPM	432746	1671565.37	919294.58	100003.00	903905.00	1661346.00	3260787.00
VIV_SEL	432746	6.89	5.69	1.00	3.00	5.00	24.00
CVE_ENT	432746	16.53	9.17	1.00	9.00	16.00	32.00
CVE_MUN	432746	39.02	62.01	1.00	7.00	19.00	570.00
HOGAR	432746	1.03	0.21	1.00	1.00	1.00	6.00
N_REN	432746	2.77	1.79	1.00	1.00	2.00	24.00
PAREN	432746	2.63	1.72	1.00	1.00	3.00	11.00
SEXO	432746	1.51	0.50	1.00	1.00	2.00	2.00
EDAD	432746	32.98	21.42	0.00	15.00	30.00	99.00
P2.5	432746	54.32	46.90	1.00	2.00	96.00	98.00
P2.6	432746	65.87	44.44	1.00	1.00	96.00	98.00
NIV	432746	3.84	2.89	0.00	2.00	3.00	11.00
GRA	416098	3.12	1.62	0.00	2.00	3.00	9.00
P2.9	416098	1.71	0.45	1.00	1.00	2.00	2.00
P2.10	416098	2.53	1.14	1.00	1.00	3.00	8.00
P2.11	416098	1.93	0.25	1.00	2.00	2.00	2.00
P2.13	351278	1.46	0.50	1.00	1.00	1.00	2.00
P2.16	351278	4.35	1.86	1.00	3.00	5.00	6.00
FAC_VIV	432746	295.70	252.31	12.00	132.00	226.00	2826.00
FAC_MUJ	432746	116.75	326.98	0.00	0.00	0.00	8836.00
ESTRATO	432746	2.14	0.82	1.00	2.00	2.00	4.00
EST_DIS	432746	298.81	176.60	1.00	154.00	300.00	617.00
UPM.DIS	432746	9276.42	5065.16	1.00	4953.00	9192.00	17795.00

Cuadro 3: Resumen estadístico de variables numéricas

4.5. Implicaciones

Los resultados confirman que:

1. Los datos faltantes en variables demográficas (EDAD) son MCAR/MAR y manejables con métodos simples.
2. Las variables sensibles (NIV) requieren enfoques avanzados (regresión) debido a su naturaleza MNAR.
3. La imputación preservó las relaciones clave entre variables, permitiendo un análisis válido de factores asociados a la violencia de género en la ENDIREH 2021.

5. Conclusión

El análisis de datos faltantes es fundamental. Los hallazgos ENDIREH 2021 revelan: tasas de no respuesta menor al 70 % en preguntas incómodas; outliers en EDAD. La imputación por media para EDAD y regresión para NIV mantienen una simetría y relaciones entre variables.

6. Bibliografía

- Aceves Fernández, M. A. (2024). *De 0 a 100 en Inteligencia Artificial*.
 Aceves Fernández, M. A. (2026). Presentaciones. Material de clase, UAQ.
 Géron, A. (2019). *Hands-On Machine Learning*. O'Reilly Media.
 INEGI. (2021). ENDIREH 2021. <https://www.inegi.org.mx/programas/endireh/2021/>