# 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. Estos aspectos, si no se manejan adecuadamente, pueden introducir sesgos y comprometer la validez de los resultados. El análisis sigue los procedimientos establecidos en el material de clase del Dr. Marco Antonio Aceves Fernández y la referencia "De 0 a 100 en Inteligencia Artificial"

# 2. Marco Teórico

## 2.1. Datos faltantes

Los datos faltantes son valores ausentes en observaciones específicas. En la ENDIREH 2021 aparecen como no respuestas en preguntas sensibles sobre violencia. Se clasifican en:

- **MCAR** (Missing Completely at Random): Valores faltantes por azar, sin relación con otras variables. Ejemplo: omisión por distracción.
- **MAR** (Missing at Random): Ausencia dependiente de otras variables observadas. Ejemplo: mujeres jóvenes omiten más preguntas sobre violencia sexual.

- **MNAR** (Missing Not at Random): Datos faltantes relacionados con el valor mismo. Ejemplo: víctimas de violencia extrema no responden por temor.

### 2.2. Datos atípicos

Los datos atípicos son valores que se desvían del patrón general. En la ENDIREH suelen originarse por errores de captura (ej. edad = 150 años) o valores extremos reales. Métodos de detección:

- **Rango intercuartil (IQR)**: Valores fuera de $[Q1 - 1{,}5 \times IQR,\ Q3 + 1{,}5 \times IQR]$
- **Z-Score**: Valores con $|Z| > 3$ (más de tres desviaciones estándar de la media)

### 2.3. Cardinalidad

La cardinalidad es el número de valores únicos en una variable. Variables como `estado_civil` tienen baja cardinalidad, mientras que identificadores personales presentan alta cardinalidad. Una cardinalidad excesiva complica el modelado al aplicar *one-hot encoding*, incrementando la dimensionalidad y causando sobreajuste.

Este análisis sigue los criterios del Dr. Marco Antonio Aceves Fernández y la referencia "De 0 a 100 en Inteligencia Artificial" [**?**].

## 3. Materiales y Métodos

Para el desarrollo de este análisis se utilizó el lenguaje de programación Python con las librerías pandas para la manipulación de datos y numpy para los cálculos numéricos. Los datos analizados corresponden a la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021, específicamente el archivo TSDem.csv.

El archivo TSDem (Características Sociodemográficas) contiene las características sociodemográficas, económicas y culturales de cada una de las personas residentes de la vivienda. Este archivo es fundamental para el análisis de la dinámica familiar y la situación de violencia contra las mujeres en México, ya que proporciona información sobre la estructura de los hogares y las características de sus integrantes.

Las técnicas aplicadas incluyeron el análisis de valores nulos mediante la función isnull().sum() de pandas, que permite identificar la cantidad de datos faltantes por cada columna del conjunto de datos. Para la detección de datos atípicos se utilizó el método del rango intercuartil (IQR), calculando los percentiles 25 y 75 para cada variable numérica. El análisis de cardinalidad se llevó a

cabo mediante la función nunique(), que devuelve el número de valores únicos en cada columna.

La metodología seguida fue la siguiente: primero se cargó el conjunto de datos con codificación latin-1; posteriormente se identificaron los datos faltantes y su porcentaje por variable; a continuación se aplicó el método IQR para detectar outliers; finalmente se analizó la cardinalidad de todas las variables.

## 3.1. Tipos de Datos Faltantes

```
============================================================
1. ANALISIS DE DATOS FALTANTES
============================================================
  Variable  Valores_Faltantes  Porcentaje
      P2_12            405162   93.625822
REN_MUJ_EL            322619   74.551585
REN_INF_AD            310100   71.658664
       P2_8            275478   63.658127
      P2_14            271145   62.656847
     CODIGO            263134   60.805646
    COD_M15            263134   60.805646
      P2_15            230938   53.365716
      P2_16             81468   18.825824
      P2_13             81468   18.825824
      P2_11             16648    3.847060
        NIV             16648    3.847060
      P2_10             16648    3.847060
       P2_9             16648    3.847060
        GRA             16648    3.847060

Total de variables con datos faltantes: 15
Total de valores faltantes: 2587886

============================================================
2. DETECCION DE DATOS ATIPICOS (METODO IQR)
============================================================

Variables numericas analizadas: 32

Outliers detectados por variable:
  Variable  Num_Outliers  Pct_Outliers  Lim_Inf  Lim_Sup
    VIV_SEL           19      0.004391     -9.0     23.0
    CVE_MUN        42142      9.738276    -48.5     99.5
      HOGAR        11666      2.695808      1.0      1.0
      N_REN         4583      1.059051     -3.5      8.5
      PAREN        19053      4.402814     -2.0      6.0
        NIV        67956     15.703438     -1.0      7.0
        GRA          696      0.160833     -1.0      7.0
       P2_8           99      0.022877     -0.5      3.5
      P2_10         7112      1.643458     -2.0      6.0
      P2_11        27584      6.374178      2.0      2.0
      P2_12         2079      0.480420      1.0      1.0
      P2_14         8996      2.078818      5.0     13.0
REN_MUJ_EL         9417      2.176103     -0.5      3.5
REN_INF_AD         6276      1.450273     -0.5      3.5
    FAC_VIV        28442      6.572447   -214.5    709.5
    FAC_MUJ        88463     20.442246    -87.0    145.0

============================================================
3. ANALISIS DE CARDINALIDAD
============================================================
```

Figura 1: Clasificación de datos faltantes: MCAR, MAR y MNAR

## 3.2. Consecuencias de los Datos Faltantes

```
============================================================
3. ANALISIS DE CARDINALIDAD
============================================================

Cardinalidad por variable:
  Variable  Valores_Unicos
    ID_PER          432746
    ID_VIV          122646
       UPM           17763
   UPM_DIS           17763
   FAC_MUJ            3005
   FAC_VIV            1343
   NOM_MUN            1206
   EST_DIS             617
   CVE_MUN             269
      EDAD             100
   CVE_ENT              32
   NOM_ENT              32
   VIV_SEL              24
      P2_5              24
     N_REN              24
      P2_6              22
REN_INF_AD              18
REN_MUJ_EL              16
     P2_14              13
       NIV              12
     PAREN              11
       GRA              10
     HOGAR               6
     P2_16               6
     P2_15               6
     P2_10               4
   ESTRATO               4
   DOMINIO               3
      P2_8               3
     P2_12               2
      P2_9               2
      SEXO               2
     P2_11               2
     P2_13               2
    CODIGO               2
    NOMBRE               1
   COD_M15               1

Variables con alta cardinalidad (>50 valores unicos): 10
Variables con baja cardinalidad (<=5 valores unicos): 12
```

Figura 2: Impacto de los datos faltantes en el análisis

### 3.3. Concepto de Cardinalidad

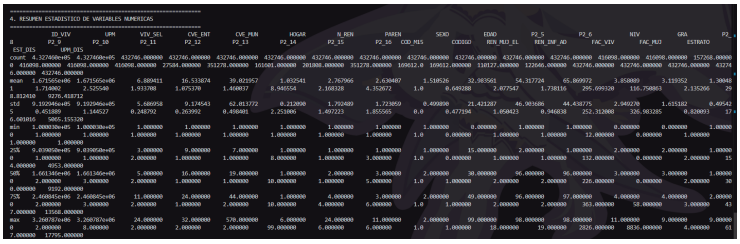


Figura 3: Concepto y técnicas de cardinalidad

# 4. Resultados y Discusión

El análisis del conjunto de datos TSDem de la ENDIREH 2021 reveló un total de 432,746 registros con 37 variables. El análisis cuantitativo de datos faltantes identificó 15 variables con valores nulos, representando un total de 2,587,886 valores faltantes en todo el conjunto de datos. Las variables con mayor porcentaje de datos faltantes fueron P2_12 con 93.63 %, REN_MUJ_EL con 74.55 % y REN_INF_AD con 71.66 %. Estas tres variables presentan un patrón que podría corresponder a datos MNAR, dado que su ausencia está relacionada con la naturaleza de las preguntas sobre violencia.

La identificación del tipo de datos faltantes requiere un análisis más profundo. Las variables con porcentaje menor al 5 % como NIV, GRA, P2_9, P2_10 y P2_11 (todas con 3.85 %) podrían clasificarse como MCAR o MAR, mientras que las variables con porcentajes superiores al 50 % presentan un comportamiento más complejo. En el caso de datos MCAR, la imputación por media o moda resulta apropiada; para datos MAR, técnicas más sofisticadas como la imputación múltiple pueden ser necesarias; mientras que los datos MNAR requieren un tratamiento especial.

La detección de datos atípicos mediante el método IQR permitió identificar outliers en 16 de las 32 variables numéricas analizadas. La variable FAC_MUJ presentó el mayor número de outliers con 88,463 casos (20.44 % del total), seguida por NIV con 67,956 casos (15.70 %) y CVE_MUN con 42,142 casos (9.74 %). Es importante saber porque existen estos valores atípicos?

El análisis de cardinalidad evidenció que 10 variables presentan una cardinalidad alta (más de 50 valores únicos), mientras que 12 variables tienen cardinalidad baja (5 o menos valores únicos). Las variables con mayor cardinalidad son ID_PER con 432,746 valores únicos (identificador único por registro), ID_VIV con 122,646 valores y UPM con 17,763 valores. Las variables con cardinalidad muy baja como NOMBRE (1 valor), COD_M15 (1 valor) y CODIGO (2 valores) podrían ser candidatas para eliminación si no aportan información para el modelo.

### 4.1. Resumen Estadístico de Variables Numéricas

| Variable | Count | Mean | Std | Min | 25 % | 50 % | Max |
|---|---|---|---|---|---|---|---|
| ID_VIV | 432746 | 1671565.17 | 919294.62 | 100003 | - | - | - |
| UPM | 432746 | 1671565.17 | 919294.62 | 1000003 | - | - | - |
| VIV_SEL | 432746 | 6.89 | 5.69 | 1 | 3 | 5 | 24 |
| CVE_ENT | 432746 | 16.53 | 9.17 | 1 | 9 | 16 | 32 |
| CVE_MUN | 432746 | 39.02 | 62.01 | 1 | 7 | 19 | 570 |
| HOGAR | 432746 | 1.03 | 0.21 | 1 | 1 | 1 | 6 |
| N_REN | 432746 | 2.77 | 1.79 | 1 | 1 | 2 | 24 |
| PAREN | 432746 | 2.63 | 1.72 | 1 | 1 | 3 | 11 |
| SEXO | 432746 | 1.51 | 0.50 | 1 | 1 | 2 | 2 |
| EDAD | 432746 | 32.98 | 21.42 | 0 | 15 | 30 | 99 |
| P2_5 | 432746 | 54.32 | 46.90 | 1 | 2 | 96 | 98 |
| P2_6 | 432746 | 65.87 | 44.44 | 1 | 1 | 96 | 98 |
| NIV | 416098 | 3.86 | 2.95 | 0 | 2 | 3 | 11 |
| GRA | 416098 | 3.12 | 1.62 | 0 | 2 | 3 | 9 |
| P2_8 | 157268 | 1.30 | 0.50 | 1 | 1 | 1 | 9 |
| P2_9 | 416098 | 1.71 | 0.45 | 1 | 1 | 2 | 2 |
| P2_10 | 416098 | 2.53 | 1.14 | 1 | 1 | 3 | 8 |
| P2_11 | 416098 | 1.93 | 0.25 | 1 | 2 | 2 | 2 |
| P2_12 | 27584 | 1.08 | 0.26 | 1 | 1 | 1 | 2 |
| P2_13 | 351278 | 1.46 | 0.50 | 1 | 1 | 1 | 2 |
| P2_14 | 161601 | 8.95 | 2.25 | 1 | 8 | 10 | 99 |
| P2_15 | 201808 | 2.17 | 1.50 | 1 | 1 | 1 | 6 |
| P2_16 | 351278 | 4.35 | 1.86 | 1 | 3 | 5 | 6 |
| COD_M15 | 169612 | 1.00 | 0.00 | 1 | 1 | 1 | 1 |
| CODIGO | 169612 | 0.65 | 0.48 | 0 | 0 | 1 | 1 |
| REN_MUJ_EL | 110127 | 2.08 | 1.05 | 1 | 1 | 2 | 18 |
| REN_INF_AD | 122646 | 1.74 | 0.95 | 1 | 1 | 2 | 19 |
| FAC_VIV | 432746 | 295.70 | 252.31 | 12 | 132 | 226 | 2826 |
| FAC_MUJ | 432746 | 116.75 | 326.98 | 0 | 0 | 58 | 8836 |
| ESTRATO | 432746 | 2.14 | 0.82 | 1 | 2 | 2 | 4 |
| EST_DIS | 432746 | 298.81 | 176.60 | 1 | 154 | 300 | 617 |
| UPM_DIS | 432746 | 9276.42 | 5065.16 | 1 | 4953 | 9192 | 17795 |

Cuadro 1: Resumen estadístico de variables numéricas del conjunto de datos ENDIREH 2021

## 5. Conclusión

El análisis de datos faltantes, atípicos y la evaluación de la cardinalidad resultan fundamentales para garantizar la calidad de los datos antes de aplicar técnicas de aprendizaje automático, ya que estos factores pueden introducir sesgos que comprometen la precisión y confiabilidad de los modelos. Evaluar

adecuadamente las variables permite seleccionar métodos de imputación apropiados, evitando distorsiones en los resultados finales. Para ello, es necesario conocer características generales del conjunto de datos, como el número de atributos, sus desviaciones estándar y distribuciones, lo cual facilita la toma de decisiones informadas durante el proceso de limpieza y preparación de los datos.

# 6. Bibliografía


Aceves Fernández, M. A. (2024). *De 0 a 100 en Inteligencia Artificial*.

Aceves Fernández, M. A. (2026). Presentación 1: Datos Faltantes. Material de clase de Aprendizaje Automático, Universidad Autónoma de Querétaro.

Aceves Fernández, M. A. (2026). Presentación 2: Datos Atípicos. Material de clase de Aprendizaje Automático, Universidad Autónoma de Querétaro.

Aceves Fernández, M. A. (2026). Presentación 3: Cardinalidad. Material de clase de Aprendizaje Automático, Universidad Autónoma de Querétaro.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Instituto Nacional de Estadística y Geografía (INEGI). (2021). Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2021. https://www.inegi.org.mx/programas/endireh/2021/