Análisis Exploratorio de Pobreza Multidimensional en México

Aplicación Modular en Python para Análisis de Base de datos y Tipo de Variables

Yhael Salvador Perez Balderas
Universidad Autónoma de Querétaro

30 de enero de 2026

1 Introducción

Este reporte presenta una aplicación modular desarrollada en Python para analizar un dataset del Instituto Nacional de Estadística y Geografía (INEGI) sobre pobreza multidimensional en México, desglosado por sexo. El objetivo es demostrar cómo una arquitectura de software bien diseñada facilita la identificación de características relevantes, la detección de valores atípicos y el balance de clases.

2 Marco Teórico

2.1 Tipos de datos en Aprendizaje Automático

Los datos utilizados en este análisis son **estructurados y tabulares**, caracterizados por observaciones (filas) y variables (columnas).A diferencia de datos no estructurados (imágenes, texto), los datos tabulares permiten un análisis estadístico directo.

2.2 Variables confusoras

En el contexto de este dataset, es relevante identificar dos tipos de variables problemáticas:

- **Variables confusoras**: Factores que influyen simultáneamente en la variable independiente y la dependiente, generando correlaciones espúreas. En este caso, el `sexo` podría actuar como variable confusora si se estudia la relación entre pobreza y acceso a servicios sin controlar por género.

- **Variables filtrantes (leaky)**: Características que contienen información comprometida con la geografía. Por ejemplo, `clave_entidad` podría ser leaky si se usa para predecir pobreza en nuevas localidades, ya que el modelo podría aprender patrones específicos de entidades conocidas en lugar de reglas generales.

3 Materiales y Métodos

3.1 Dataset analizado

Se utilizó el archivo `pobreza_grupos_poblacionales_sexo.csv` del INEGI, que contiene información a nivel de localidad sobre indicadores de pobreza multidimensional. Cada localidad aparece **dos veces** en el dataset: una para la población femenina y otra para la masculina. Esto permite comparar cómo las condiciones de pobreza afectan a ambos géneros dentro del mismo contexto geográfico.

Cuadro 1: Estructura del dataset

Tipo de variable	**Ejemplos en el dataset**
Identificadores geográficos	`clave_entidad`, `clave_municipio`
Demográficas	`poblacion`, `grupo` (sexo)
Indicadores de pobreza	`pobreza_porcentaje`, `carencia_rezago_educativo_porcentaje`, `ingreso_inferior_a_lpi_porcentaje`

El dataset original contiene 14,764 registros (7,382 localidades $\times$ 2 sexos) y 17 atributos. Los valores faltantes están codificados como `-999.0` y fueron convertidos a `NaN` durante la carga.

3.2 Arquitectura de la aplicación

La aplicación sigue el principio de **separación de responsabilidades**, organizada en tres módulos:

- `core/`: Contiene la lógica de negocio.
  - `data_loader.py`: Carga el CSV y gestiona valores faltantes mediante `na_values=[-999.0]` en `pandas.read_csv()`.
  - `analyzer.py`: Calcula estadísticas descriptivas (mínimo, máximo, desviación estándar) y balance de clases.
- `ui/`: Gestiona la interacción con el usuario.
  - `console.py`: Solicita rutas y muestra resultados en consola.

- `visualizer.py`: Genera histogramas comparativos por sexo.

- `main.py`: Orquesta el flujo completo:

  1. Solicita la ruta del archivo CSV al usuario.
  2. Carga y preprocesa los datos (extracción de sexo desde `grupo`).
  3. Ejecuta análisis estadísticos.
  4. Muestra resultados y genera visualizaciones.

4 Resultados y Discusión

4.1 Resumen estadístico

El análisis arrojó los siguientes resultados clave:

Cuadro 2: Estadísticas descriptivas de variables seleccionadas

Variable	**Mínimo**	**Máximo**	**Desv. Estándar**
Población	38.0	999,227.0	71,284.09
Pobreza (%)	3.33	99.95	20.76
Rezago educativo (%)	1.74	67.70	10.76
Servicios básicos (%)	0.48	100.00	29.10
Ingreso inferior a LPI (%)	4.45	99.95	18.53

- La **población** presenta alta variabilidad (desv. estándar: 71,284), reflejando la coexistencia de localidades rurales pequeñas y zonas urbanas densas.

- El rango de **pobreza** (3.33 %–99.95 %) evidencia una distribución heterogénea: algunas localidades tienen mínima incidencia de pobreza, mientras que otras presentan casi totalidad de su población en condición de pobreza.

- La **desviación estándar más alta** corresponde a *servicios básicos* (29.10), indicando gran disparidad en el acceso a agua potable, drenaje y electricidad entre localidades.

4.2 Balance de clases

El dataset muestra un balance perfecto entre géneros:

- Hombres: 7,382 registros (50.00 %)

- Mujeres: 7,382 registros (50.00 %)

Este equilibrio valida la estructura del dataset (doble registro por localidad) y elimina sesgos muestrales en análisis comparativos por sexo.

4.3 Visualización: Histogramas por sexo

La Figura 1 muestra la distribución de indicadores clave separados por sexo. Los histogramas revelan:

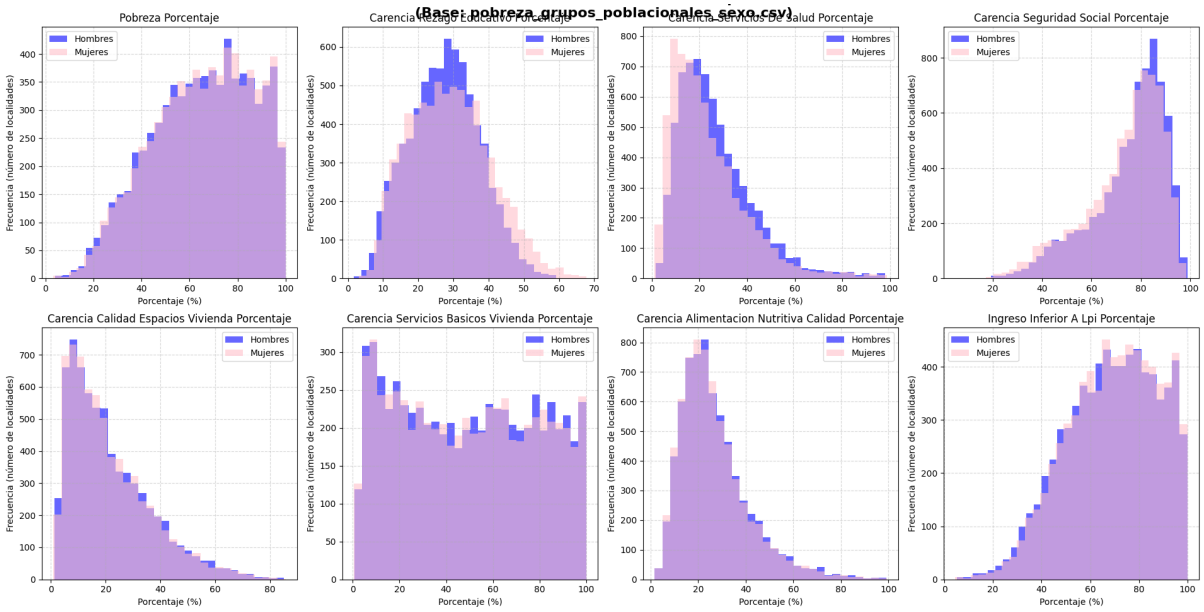

Figura 1: Distribución de indicadores de pobreza por sexo. Azul: hombres; Rosa: mujeres.

- **Distribuciones similares**: Para la mayoría de indicadores (pobreza, rezago educativo), las curvas de hombres y mujeres son casi idénticas, sugiriendo que la pobreza afecta a ambos géneros de manera comparable a nivel agregado.
- **Simetría en ingresos**: La distribución de *ingreso inferior a LPI* es prácticamente idéntica para ambos sexos, reflejando que la pobreza extrema por ingreso no discrimina por género a nivel municipal.

Estos hallazgos son valiosos para el diseño de políticas públicas con enfoque de género, ya que identifican dimensiones específicas donde las mujeres podrían requerir atención prioritaria.

5 Conclusión

La aplicación desarrollada demuestra cómo una arquitectura modular en Python facilita el análisis exploratorio de datos estructurados, paso previo indispensable en cualquier proyecto de Aprendizaje Automático. Los resultados obtenidos del dataset del INEGI revelan:

- Una distribución heterogénea de la pobreza multidimensional en México, con localidades que van desde mínima incidencia hasta casi totalidad de su población en condición de pobreza.

- Similitudes generales entre géneros en la mayoría de indicadores, pero diferencias sutiles en dimensiones específicas como servicios básicos.
- La importancia de identificar variables confusoras (sexo) y filtrantes (clave_entidad) antes de entrenar modelos predictivos.

Se podrían predecir niveles de pobreza a partir de características demográficas y geográficas, siempre evitando el uso de variables filtrantes para garantizar la generalización del modelo.

6 Bibliografía

1. INEGI. (2023). *Microdatos de pobreza multidimensional.*
2. CONEVAL. (2022). *Guía para la medición de la pobreza multidimensional en México.* Ciudad de México.