UNIVERSIDAD DE BUENOS AIRES - 2025

Big Data y Machine Learning

Trabajo Práctico N° 2

Un primer encuentro con la EPH

Fecha de entrega: 23 de septiembre a las 13:00 hs.

Contenido*:* familiarización con la base de datos de la Encuesta Permanente de Hogares. Limpieza de datos, valores faltantes y análisis descriptivo. Medición de pobreza.

Integrantes:

- Antúnez, María (
- Pistoia, Stefano (
- Signorelli, Franco (895391)

Link al Repositorio del grupo:
https://github.com/StefanoPistoia/Big-Data-Grupo-2/tree/main

Parte I: Familiarizandonos con la base EPH y limpieza

1. *"Utilizando información disponible en la página del INDEC, expliquen brevemente cómo se identifica a las personas pobres."*

El criterio vigente establecido por el INDEC para la identificación de personas pobres en Argentina se define a partir de un indicador general denominado *línea de pobreza* (en adelante, "LP"), el cual tiene una estrecha relación con otro llamado *línea de indigencia* (en adelante, "LI").

La línea de indigencia es un método de medición indirecta que "*procura establecer si los hogares cuentan con ingresos suficientes para cubrir una canasta de alimentos capaz de satisfacer un umbral mínimo de necesidades energéticas y proteicas, denominada Canasta Básica Alimentaria (CBA)"* (INDEC, 2016).

La CBA es una canasta *representativa* a la cual se le asigna un índice y cuya evolución responde a la variación de los precios al consumidor de los alimentos que la componen. Por tanto, se define con el carácter de *indigente* a aquel hogar o individuo que no cuente con los ingresos suficientes para cubrir el valor de la canasta.

Adicionalmente, la LP funciona como una extensión de la LI, en el sentido de que evalúa si los ingresos de los individuos u hogares permiten cubrir no sólo la Canasta Básica Alimentaria, sino también una canasta que contemple *"otros consumos básicos no alimentarios"* (INDEC, 2016), denominada Canasta Básica Total (en adelante, "CBT").

La relación cuantitativa para la medición de estos índices es la siguiente:

$$CBT = CBA * \text{Gasto Alimentario} / \text{Gasto Total}$$

De esto se desprende que esta metodología de medición fija al gasto alimenticio como una proporción constante del gasto total.

2. Elección de la región de análisis y variables de interés; comentarios sobre el proceso de limpieza de las bases de datos

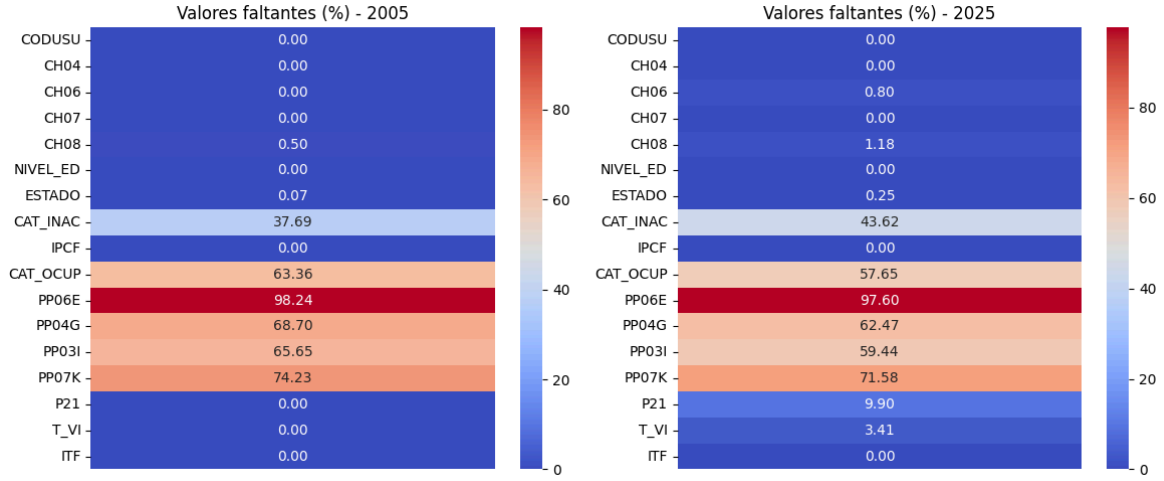
    a. Región de interés

Hemos elegido el Noreste (NEA), con '41' como su código asociado en la EPH.

    b. Variables de interés elegidas

- CAT_OCUP: Categoría ocupacional
- PP06E: Tipo de constitución jurídica de la actividad de trabajadores independientes (S.A., S.R.L., sociedad de palabra, etc.)
- PP04G: ¿Dónde realiza principalmente sus tareas?
- PP03I: ¿En los últimos 30 días, buscó trabajar más horas?
- PP07K: Documentación asociada al cobro de los asalariados
- P21: Monto de ingreso de la ocupación principal
- T_VI: Monto de ingresos no laborales

El *heatmap* de valors faltantes resultante es el siguiente:



Fuente: elaboración propia en base a EPH (INDEC)

Del cual se desprende fácilmente que 'PPO06E' (Tipo de constitución jurídica de la actividad de trabajadores independientes) es, en ambas encuestas, la variable con menor cantidad de registros (con sólo un 2-3% de respuestas).

El resto de las variables relacionadas a las características de la actividad ocupacional de cada encuestado presentan *missing values* en un rango que oscila entre el 60% y el 75% si se tienen en cuenta ambas encuestas.

c. Limpieza de datos

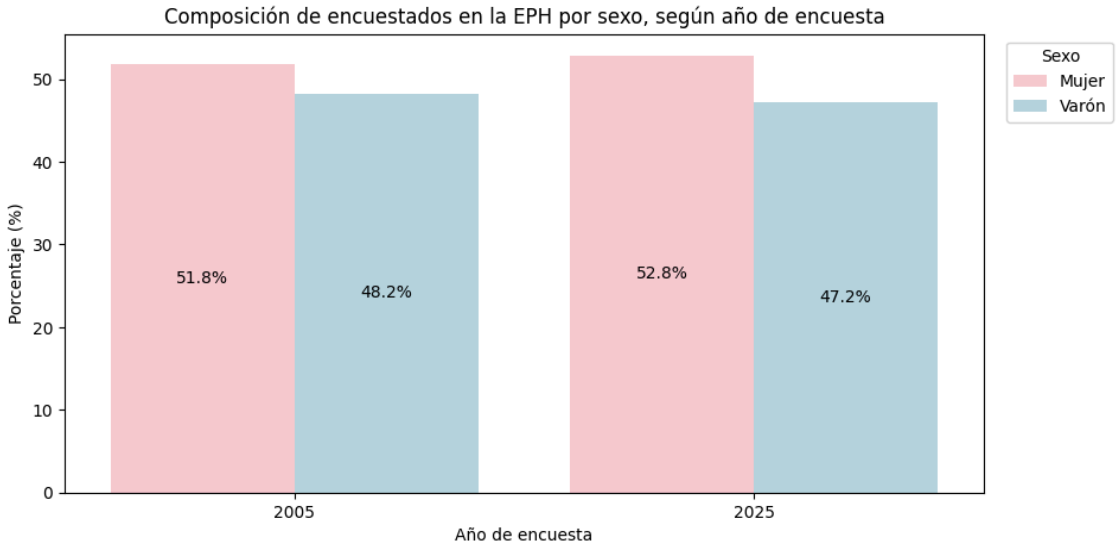
El primer criterio de elección de variables que elegimos fue que las variables estuvieran presentes en ambas encuestas. Por tal motivo, nuestra primer etapa de limpieza fue identificar qué columnas se presentaban solo en uno de los dos datasets y descartarlas. Ya elegidas 7 variables que cumplieran con esta condición, elaboramos un dataframe concatenando ambas encuestas y conteniendo sólo las 15 variables de interés.

El siguiente paso fue de homogeneización de los datos. Para ello, construimos una secuencia de funciones para identificar los valores únicos por columna, modificar algunos de ellos, y convertir el *datatype* de las columnas a *float* con el

doble objetivo de facilitar el procesamiento de datos y alivianar el peso del dataframe.

Parte II: Primer Análisis Exploratorio

3. "Realicen un gráfico de barras mostrando la composición por sexo para 2005 y 2025 en su región. Comenten los resultados."



Fuente: elaboración propia en base EPH(INDEC)

Analizando la composición de encuestados por sexo, tenemos un claro ejemplo de cómo las muestras de datos, cuando son bien utilizadas, pueden dar información confiable de los datos poblacionales. Es sabido que la distribución poblacional por sexo evaluada en casi toda dimensión social es aproximadamente la que arroja como resultado esta encuesta: "un poco más" de la mitad son mujeres, y "un poco menos" de la mitad son varones.

4. "Realicen una matriz de correlación para 2005 y 2025 con las siguientes variables: CH04, CH06, CH07, CH08, NIVEL ED, ESTADO, CAT_INAC, IPCF. Crear las variables dicotómicas binarias necesarias (variables dummies) y renombrar dichas variables para que las etiquetas tengan sentido en el gráfico de correlación. Utilicen alguno de los comandos disponibles en este link para graficar la matriz de correlación. Comenten los resultados."

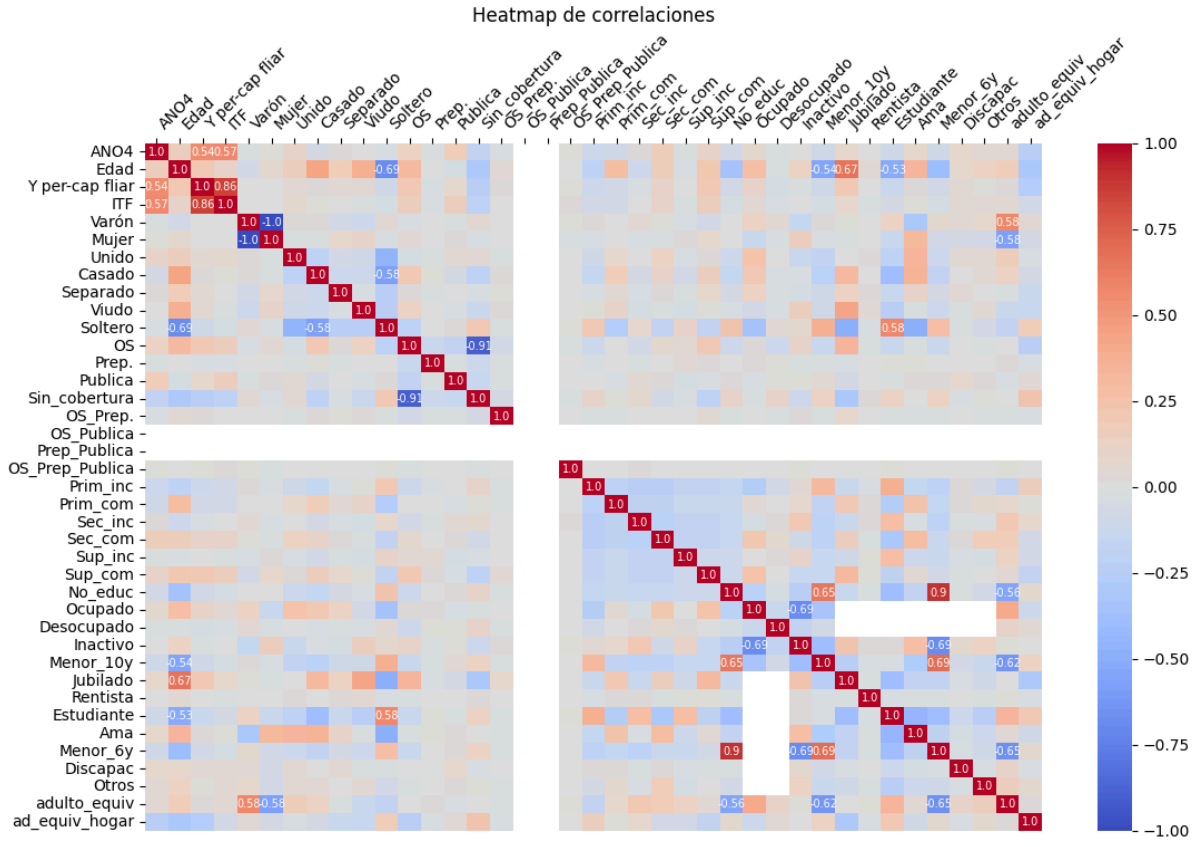
El trabajo sobre el dataframe para el posterior *heatmap* de correlaciones consistió en desdoblar cada una de las variables categóricas en una columna por cada categoría eixstente al interior de ellas.

Así, por ejemplo, la variable 'CH08' (*Tipo de cobertura médica)*, que presentaba 9 categorías:

- 1 = Obra social (incluye PAMI)
- 2 = Mutual/prepaga/servicio de emergencia
- 3 = Planes y seguros públicos
- 4 = No paga ni le descuentan
- 9 = Ns/Nr
- 12 = Obra social y mutual/prepaga/servicio de emergencia
- 13 = Obra social y planes y seguros públicos
- 23 = Mutual/prepaga/servicio de emergencia y planes y seguros públicos
- 123 = Obra social y mutual/prepaga/servicio de emergencia y planes y
- seguros públicos

Fue desdoblada en 8 variables dicotómicas (excluyendo 'Ns/Nr' por haber reemplazado dichas entradas por *NaN's*).

El *correlation heatmap output* es el que se muestra a continuación:



Fuente: elaboración propia en base a EPH (INDEC)

En él, hemos destacado las correlaciones con valor absoluto mayor a 0,5.

Algunas conclusiones que se desprenden de su análisis son:

I. Hay correlación perfectamente inversa entre varones y mujeres
→ evidencia de buen procesamiento de datos, sabiendo que estos dos campos no contienen *missing values*.

II. Hay fuerte correlación (0,86) entre el Ingreso Total Familiar ('ITF') y el Ingreso Per Cápita Familiar ('IPCF').
→ sabiendo que por definición $IPCF = ITF / (integrantes\ del\ hogar)$; es decir, que hay colinealidad entre ambas variables, la correlación en el *heatmap* no es perfecta porque la variable 'integrantes del hogar' no entra en las dimensiones de análisis de este dataframe.

III. Hay una correlación inversa considerable entre la edad y los status de *soltero/a* (-0,69) y *estudiante* (-0,53).

IV. La desocupación no guarda correlación observable con ninguna otra variable de análisis.
→ ¿Significa que es estructural, entendiéndose esto como transversal a cualquier segmentación demográfica que podamos hacer?

V. La categoría ocupacional de *ama de casa* guarda cierta correlación directa (aunque menor a 0,5) con los status de *mujer, unida* y *casada*; lo cual otorga evidencia en apoyo al concepto de la *familiarización del cuidado del hogar*. Una autora referente en esta materia es Valeria Esquivel[1].

Parte III: Conociendo a los pobres y no pobres

5. "¿Cuántas personas no respondieron cuál es su condición de actividad?"

Las cifras de no-respuesta en ambas encuesta fueron las siguiente:

En términos absolutos

Año de encuesta	2005	2025
Resultado		
Con rta.	5426	3340
Sin rta.	27	1406
Total	5453	4746

En términos relativos

Año de encuesta	2005	2025
Distribución (%)		
Con rta.	99.5	70.4
Sin rta.	0.5	29.6
Total	100.0	100.0

Se evidencia que de un grado de respuesta de 99% en 2005 se descendió a un 70% en 2025 (siendo incluso esta última cifra suavizada por métodos de asignación de valores).

8. "Por último, agreguen a `respondieron` una columna llamada `pobre` que tome valor 1 si el ITF es menor al ingreso necesario que necesita esa familia, y 0 en caso contrario. ¿Cuántos pobres identificaron para cada año? ¿Qué porcentaje de la muestra representa?"

[1] Ver: https://bicyt.conicet.gov.ar/fichas/p/valeria-renata-esquivel

A partir de la contrastación del Ingreso Total Familiar versus la CBT, las cifras de *línea de pobreza* para cada año de encuesta son las siguientes:

En términos absolutos			En términos relativos		
Distribución (%)	No pobres	Pobres	Distribución (%)	No pobres	Pobres
Año de encuesta			Año de encuesta		
2005.0	2530	2896	2005.0	46.6	53.4
2025.0	1664	1676	2025.0	49.8	50.2

9. "Muestren estadísticas descriptivas relevantes de pobre en una tabla, comparando 2005 con 2025. Además, hagan 2 gráficos exploratorios a elección usando la variable pobre. Comenten."

Optamos por analizar descriptiva la pobreza según las distintas categorías de interés, descriptas a continuación:

- Pobreza por Sexo

		Varón	Mujer
ANO4	**pobre**		
2005.0	**no_pobre**	46.10	47.12
	pobre	53.90	52.88
2025.0	**no_pobre**	48.75	50.76
	pobre	51.25	49.24

- Pobreza por Estado civil

		Unido	Casado	Separado	Viudo	Soltero
ANO4	**pobre**					
2005.0	**no_pobre**	39.62	60.79	55.61	67.53	40.04
	pobre	60.38	39.21	44.39	32.47	59.96
2025.0	**no_pobre**	46.37	63.38	59.24	71.61	43.51
	pobre	53.63	36.62	40.76	28.39	56.49

- Pobreza por Nivel educativo

ANO4	pobre	Prim_inc	Prim_com	Sec_inc	Sec_com	Sup_inc	Sup_com	No_educ
2005.0	no_pobre	30.29	40.33	40.06	63.54	66.56	87.78	35.11
	pobre	69.71	59.67	59.94	36.46	33.44	12.22	64.89
2025.0	no_pobre	37.38	53.52	33.18	52.40	58.36	81.17	35.22
	pobre	62.62	46.48	66.82	47.60	41.64	18.83	64.78

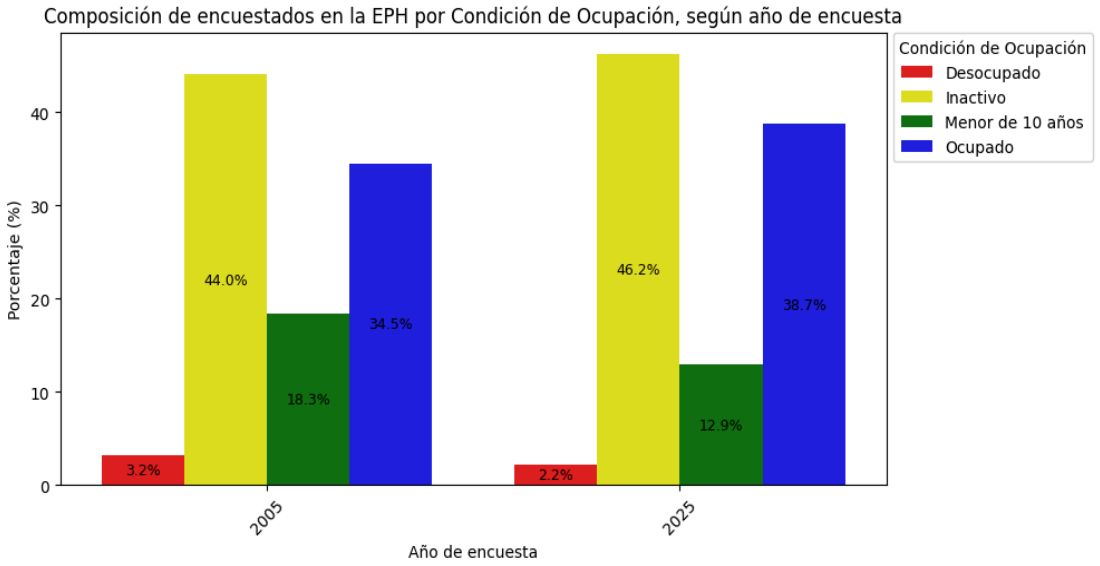
Resulta llamativo que para todos los niveles aumentó la pobreza relativa, a excepción del nivel más bajo de todos, *primaria incompleta*, aunque es posible que esto guarde relación con una disminución de la pobreza infantil.

- Pobreza por Estado de inactividad

ANO4	pobre	Jubilado	Rentista	Estudiante	Ama	Menor_6y	Discapac	Otros
2005.0	no_pobre	75.58	66.67	37.77	43.90	35.98	20.00	31.41
	pobre	24.42	33.33	62.23	56.10	64.02	80.00	68.59
2025.0	no_pobre	89.03	50.00	32.57	40.45	32.56	60.47	46.03
	pobre	10.97	50.00	67.43	59.55	67.44	39.53	53.97

Estas son las dos dimensiones que hemos elegido analizar gráficamente:

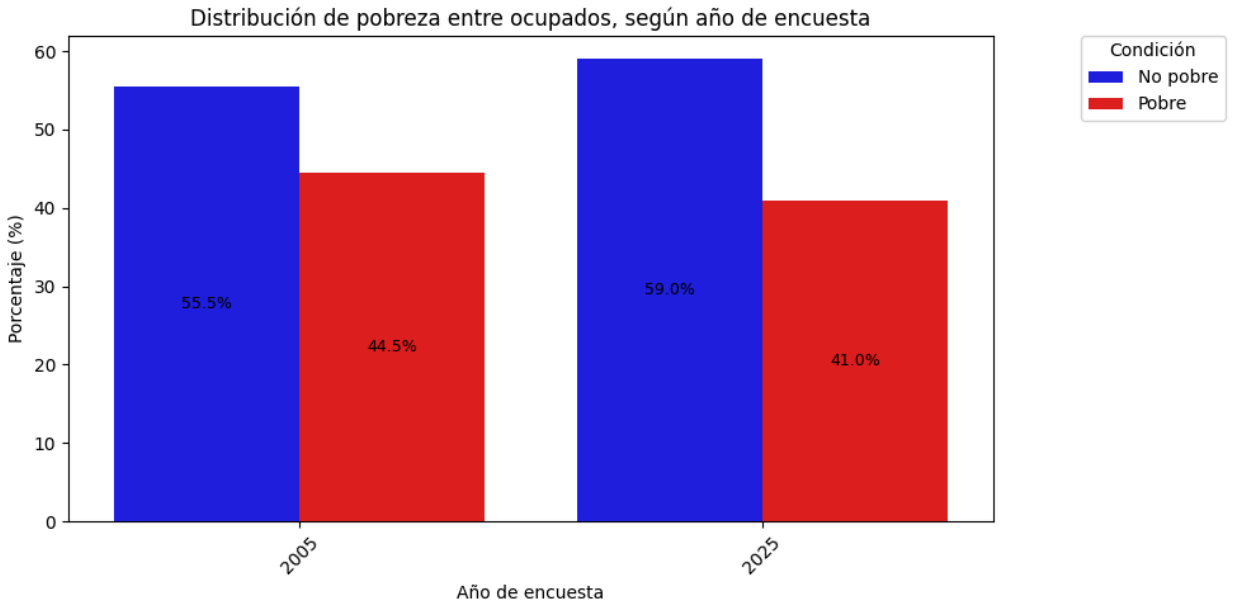
I. Composición de la *condición de ocupación:*



Algunas conclusiones a partir del análisis de este gráfico son:

- Evidencia de la regresión de pirámide poblacional: aumento de los *inactivos* (en línea con una ligera mayor esperanza de vida) y disminución de los *menores de 10 años* (baja de la tasa de natalidad)

II. Dentro de los *ocupados, medición de los ocupados pobrezs*



Algunas conclusiones a partir del análisis de este gráfico son:

- Disminución relativa de los trabajadores pobres en el Noreste Argentino. Bajo esta dimensión, no resulta posible esbozar causas y/o diferencias entre factores regionales (diferencias entre población y pobreza de las provincias que la componen) y nacionales (variación de pobreza alineada con una tendencia nacional).

Bibliografía

- Medición de pobreza: https://www.indec.gob.ar/ftp/cuadros/sociedad/EPH_metodologia_22_pobreza.pdf