TRABAJO PRÁCTICO N° 2

Histogramas, Kernels & Métodos no supervisados usando la EPH

Ricardo Javier Gutiérrez Vistín - Martín Gabriel Cargnel

Link al repositorio: https://github.com/mcargnel/Big-Data-UBA-Grupo-09/tree/main/TP2

1 Creación de variables, histogramas, kernels y resumen de la base de datos final

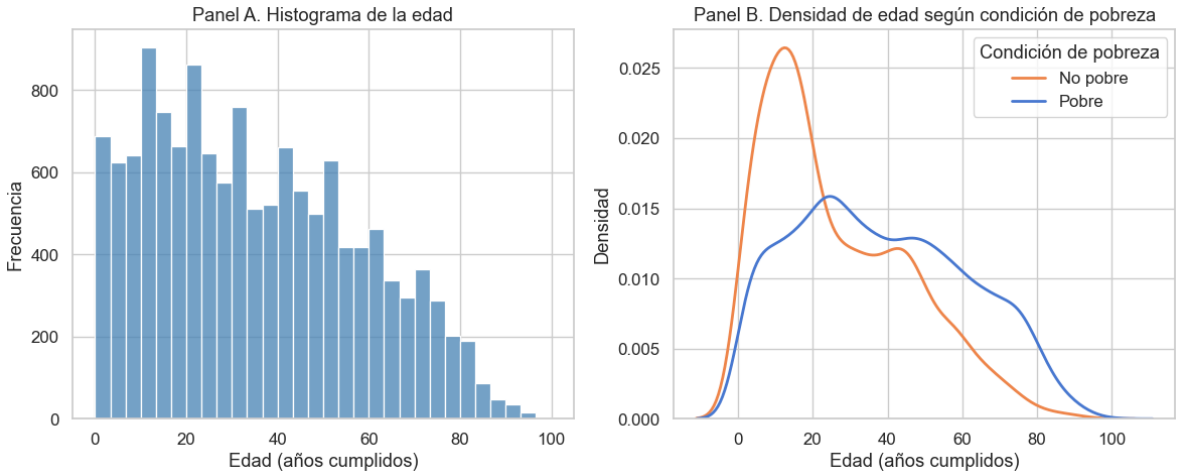


Figura 1

La distribución de edades muestra una concentración importante en los grupos jóvenes, principalmente entre los 10 y 30 años, que representan alrededor de un tercio de la población analizada, mientras que la frecuencia disminuye progresivamente a medida que aumenta la edad. En términos de pobreza, los datos indican que la incidencia es mayor en los grupos etarios más jóvenes con más del 40% de personas pobres entre 0 y 30 años y tiende a reducirse en los grupos de mayor edad, donde las tasas descienden por debajo del 15%. Esto sugiere que la vulnerabilidad económica afecta con mayor intensidad a la población joven, posiblemente debido a su menor estabilidad laboral y menor acumulación de capital humano.

	Estadístico	Valor
0	Promedio	10.79
1	Desviación estándar	4.56
2	Mínimo	0.00
3	Mediana (p50)	12.00
4	Máximo	24.00

La distribución del ingreso total familiar se concentra fuertemente en los niveles bajos, ya que casi el 98% de los hogares tiene ingresos inferiores a 9,3 millones de pesos de 2025, mostrando una clara asimetría hacia la derecha. En cuanto a la condición de pobreza, se observa que los hogares pobres se concentran principalmente en los tramos de menor ingreso, mientras que en los niveles altos prácticamente no existen, reflejando una marcada desigualdad en la distribución del ingreso.

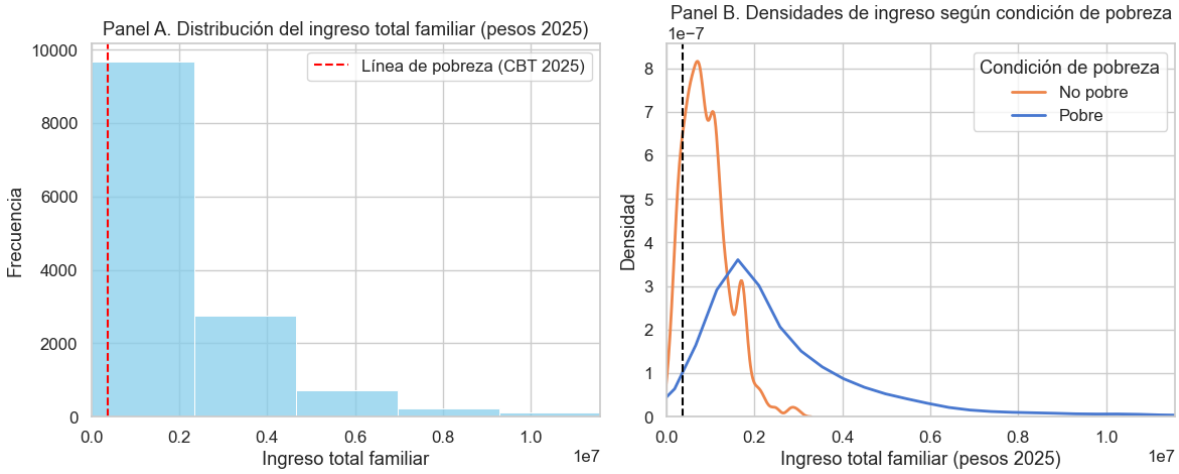


Figura 2: Visualización de la distribución de ingresos.

Tabla 2: Tabla con estadísticos sobre las horas trabajadas.

	Estadístico	Valor
0	Promedio	43.18
1	Mediana	44.00
2	Desviación estándar	18.90
3	Mínimo	1.00
4	Máximo	90.00

Tabla 3: Tabla 1. Resumen de la base final para la región seleccionada.

Año	2005	2025	Total
Cantidad de observaciones	9371	4282	13653
Observaciones con NA en 'pobre'	0	0	0
Cantidad de pobres	2493	1462	3955
Cantidad de no pobres	6878	2820	9698
Cantidad de variables limpias/homogeneizadas	34	34	34

2 Métodos no supervisados

En la Figura 3 se puede ver la correlacion de las variables edad, edad al cuadrado, años de educación, ingreso_total_familiar, el número de miembros en el hogar y horas trabajadas. La única variable que tiene una alta correlación es edad con edad al cuadrado, lo cual era

esperable. Con estas variables se implementarán métodos no supervisados para intentar dividir nuestra muestra entre pobres y no pobres. Cabe aclarar que previo a correr estos algoritmos se estandarizaron las variables dado que son sensibles a las escalas en las que están los datos.

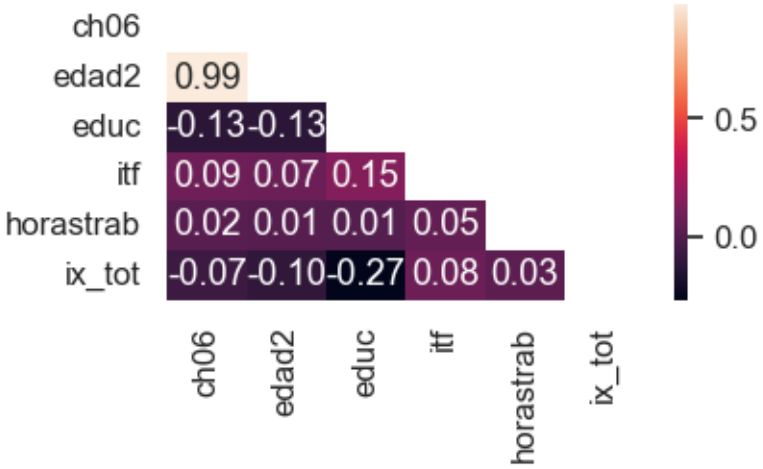


Figura 3: Heatmap con correlaciones de las variables incluidas.

2.1 PCA

En la Figura 4 vemos los scores en el panel A y los ponderadores en el B. En éste último se ve que la edad y la edad al cuadrado tienen una ponderación más alta en el primer componente, mientras que la educación y la cantidad de personas en el hogar dominan en el segundo.

En la Figura 5 se puede ver como el primer y segundo componente capturan el 55% de la variabilidad y dado que el tercero y cuarto capturan 0.18 y 0.16, respectivamente, considerando los 4 tendriamos un 89%. El quinto componente aumenta la varaibilidad explicada pero menos, siendo solo un 10%.

2.2 Cluster

En la Figura 6 se pueden ver los resultados del algoritmo de cluster k-medias donde el mismo logra identificar pobres y no pobres con k=2.

Como se observa en la Figura 1, el método del codo (o elbow) no arroja un resultado concluyente, ya que no se aprecia una reducción abrupta de la inertia para ningún número de clústeres. Esto contrasta con la expectativa de encontrar un punto de inflexión en k=2, que correspondería a los dos grupos teóricos de interés (pobres y no pobres).

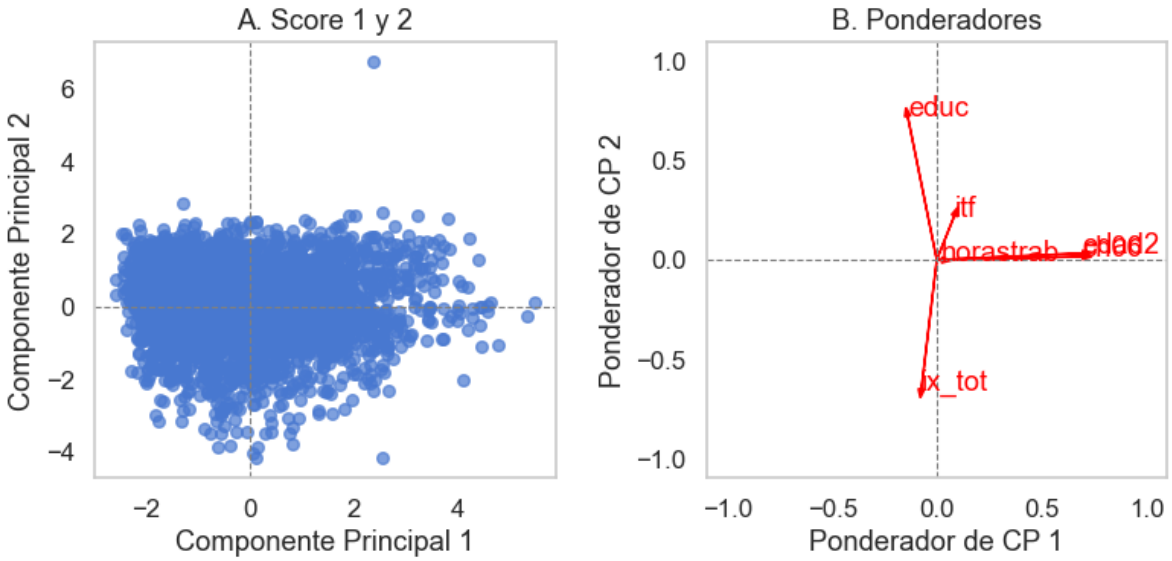


Figura 4: Visualización los scores y ponderadores.

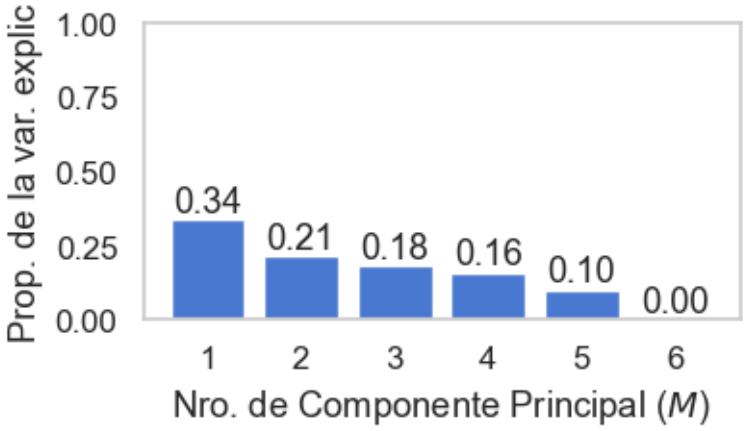


Figura 5: Proporción de la variabilidad explicada por cada componente.

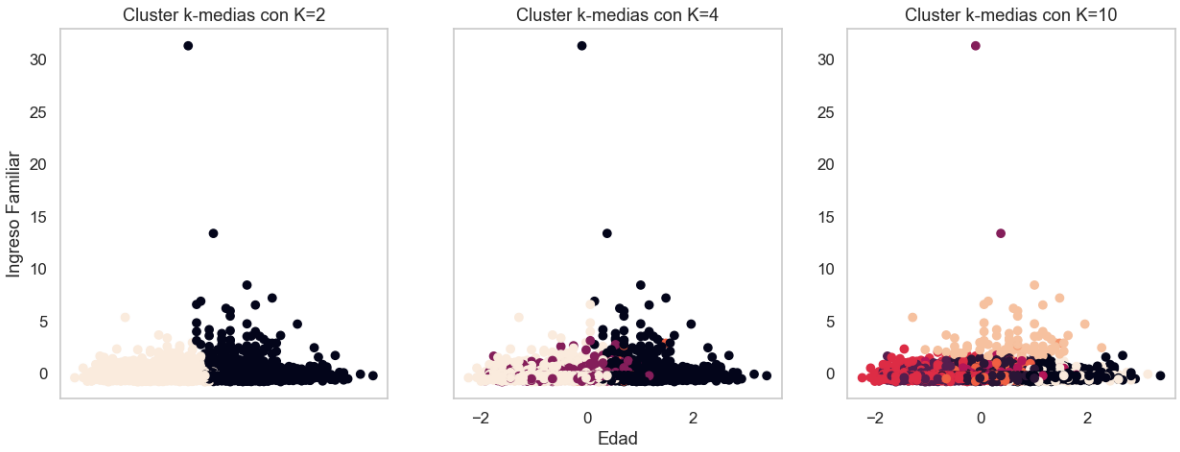


Figura 6: Cluster con distintos k.

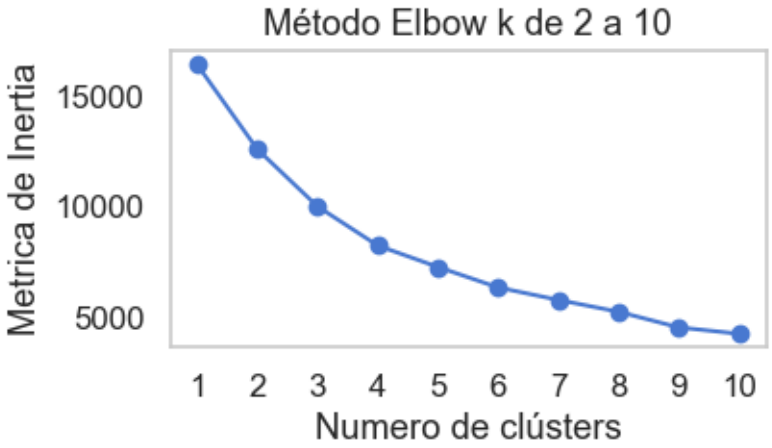
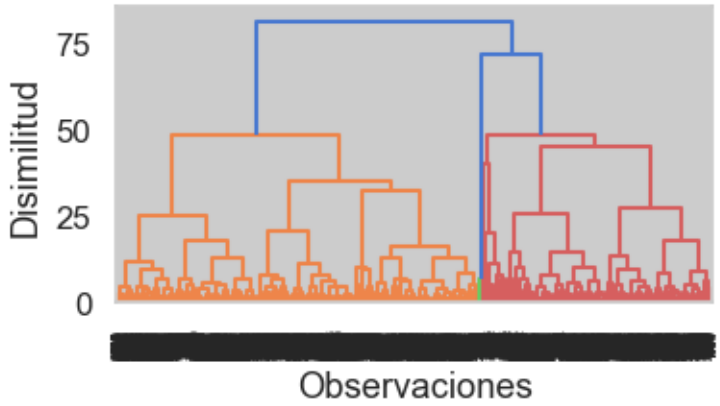


Figura 7: Elbow plot para k de 2 a 10.

Finalmente, en la Figura 8 se presenta un dendrograma con las variables utilizadas. Este permite visualizar dos particiones principales, que podrían interpretarse como pobres y no pobres. El dendrograma es un diagrama que muestra cómo se agruparon los datos en función de su distancia, de modo que aquellos que aparecen juntos son más parecidos entre sí.

```
Text(0, 0.5, 'Disimilitud')
```

(a) Dendograma.



(b)

Figura 8