Imagen que contiene Texto

Descripción generada automáticamente

Universidad de Buenos Aires.

Facultad de Ciencias Económicas

Escuela de Estudios de Posgrado.

**MAESTRÍA EN ECONOMÍA APLICADA**

Taller de Programación

Segundo Trabajo Práctico:

**Histogramas, Kernels & Métodos No Supervisados usando la EPH**

**Autores:**

Kostzer Federico

Meza Palma Manuel H.

Pardini Maximiliano G.

**docente:** Romero Maria Noelia

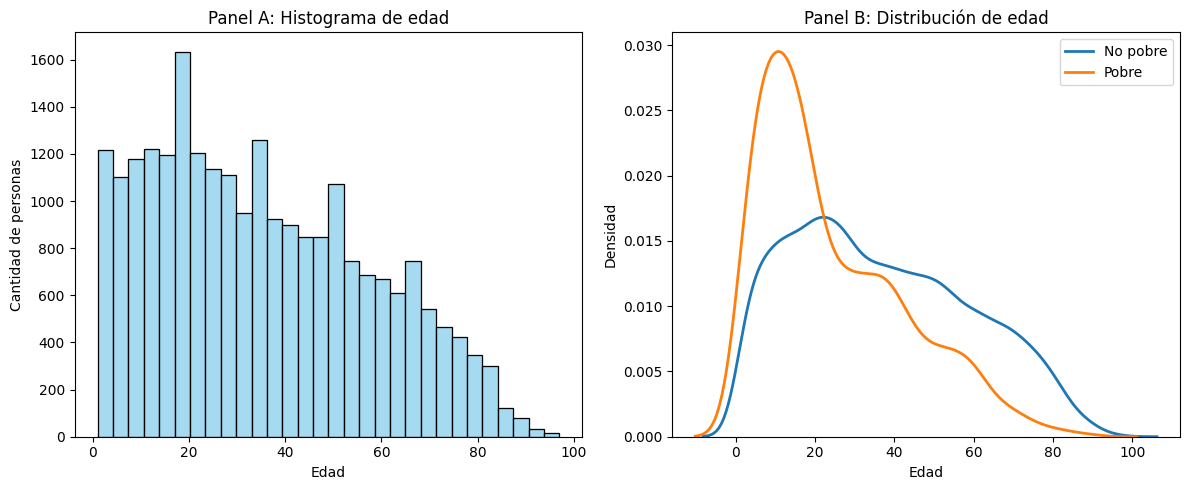
Octubre 2025

**Parte I: Creación de variables, histogramas, kernels y resumen de la base de datos final**

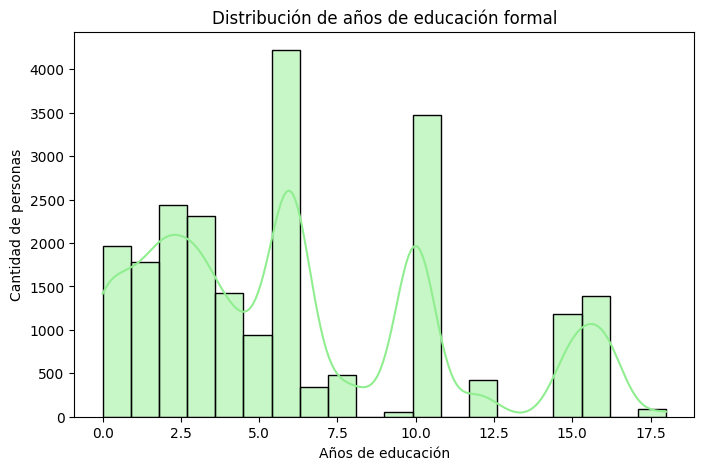
1. **Distribución de la edad por condición de pobreza (Panel A y Panel B)**

El análisis descriptivo de la variable edad (ch06) muestra que la muestra está compuesta por 23.569 observaciones, con una edad promedio de 32,5 años y una desviación estándar de 22 años, lo que evidencia una distribución bastante amplia. La edad mínima registrada es de 1 año y la máxima de 97 años. Al desagregar por condición de pobreza, se observa que los hogares no pobres presentan una edad promedio de 36,7 años, mientras que los hogares pobres tienen una media significativamente menor, de 24,3 años, reflejando una población más joven.

En cuanto a los gráficos, el histograma muestra una mayor concentración de individuos entre los 20 y 40 años, con una caída casi constante hacia edades avanzadas en torno a los 50 aaños. El gráfico de densidad (KDE), por su parte, evidencia que la distribución etaria de los hogares pobres está desplazada hacia la izquierda con una fuerte concentración alrededor de los 15 años, indicando que la pobreza afecta con mayor intensidad a grupos etarios más jóvenes.

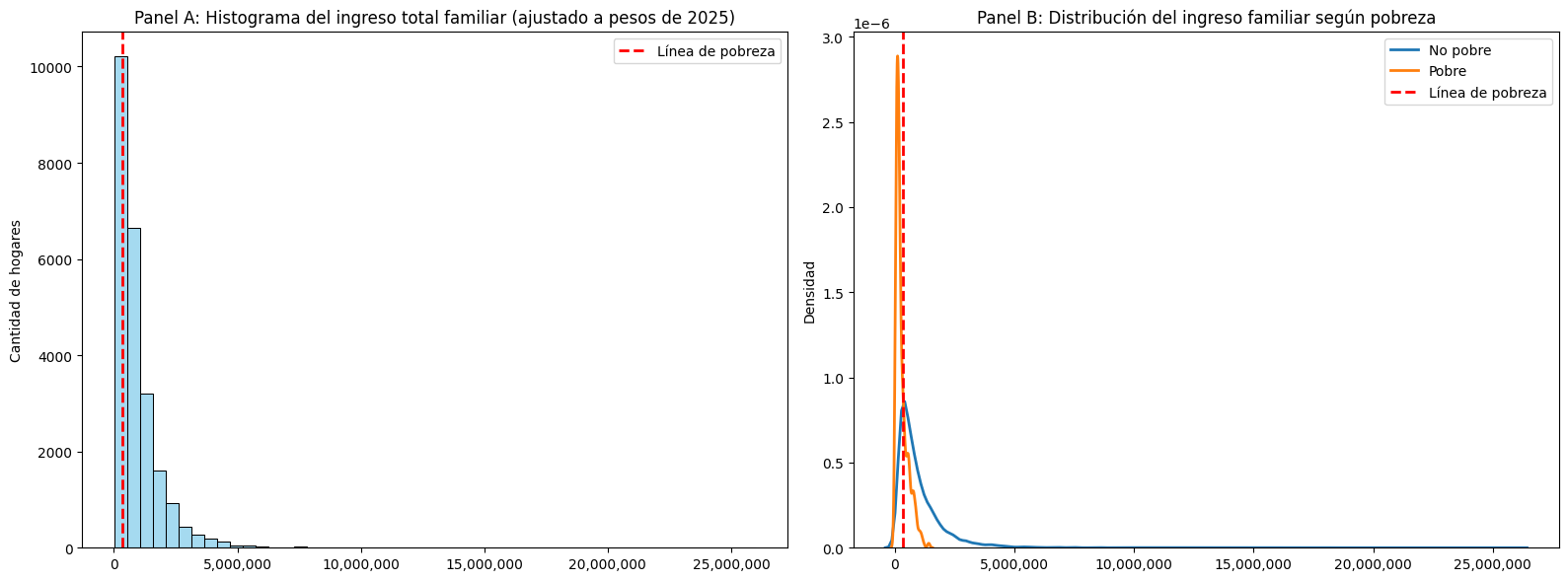


1. **Distribución de los años de educación formal**

La variable de años de educación (educ), que mide la cantidad de años de educación formal, presenta un promedio de 6 años con una desviación estándar de aproximadamente 4,72, lo que indica una alta dispersión en el nivel educativo alcanzado entre los individuos. El valor mínimo es 0 años (personas sin educación formal) y el máximo es 18 años, correspondiente a quienes completaron estudios de posgrado. La mediana (p50) coincide con la media en 6 años, lo que sugiere que al menos la mitad de los encuestados completó la educación primaria, esta alineación también indica cierta simetría entre los valores alrededor de la media. En el histograma, la distribución se concentra en los primeros años de escolaridad, con una caída progresiva hacia los niveles educativos más altos mostrando fuertes concentraciones en la media y alrededor de los 10 años de estudio, reflejando un patrón de baja escolarización promedio en la muestra analizada.

1. **Ingreso familiar**

La distribución del ingreso total familiar ajustado a 2025 muestra una alta dispersión, con un promedio de 926.482 pesos, un valor mínimo de 8.496 y un máximo de 26.000.000 pesos, lo que evidencia la existencia de hogares con ingresos muy heterogéneos. En el Panel A, el histograma refleja que la mayoría de los hogares perciben un ingreso menor a los 5 millones, pero ubicándose la gran mayoría por encima de la línea de la pobreza. En el Panel B, las distribuciones kernel muestran que los hogares pobres se concentran en valores bajos de ingreso, por debajo de la línea de pobreza, mientras que los hogares no pobres presentan una distribución más amplia y desplazada hacia ingresos mayores.



Coeficiente de actualización 2005–2025 obtenido multiplicando los factores de inflación oficial anual y mensual reportados por el INDEC y complementados con fuentes confiables (lanacion.com.ar, infobae.com, diariodemocracia.com, diarioc.com.ar, mdzol.com, pagina12.com, ámbito.com, chequeado.com).

1. **Distribución de horas trabajadas por el jefe del hogar**

En un primer ejercicio os jefes de hogar trabajaron en promedio 15 horas con una desviación típica de 32 horas, sin embargo, se reportaban jornadas de mas de cien horas y existían valores atípicos como 999 lo que introducía problemas de interpretación en los resultados, se acotaron los valores a menos de cien horas trabajadas tomando la media encontrada en la base original mas tres desvíos estándar valor que sigue siendo alto, pero más plausible.

Tras realizar nuevamente las estimaciones se obtuvo una media de 14.7 horas semanales y una media de cero horas; se redujo la dispersión de los datos ubicándose la desviación estándar en 22.2 horas, un problema que permaneció insoluble fue la gran cantidad de personas que reportaron haber trabajado cero horas durante la semana y para las cuales no había un criterio de exclusión para proceder a la limpieza lo que impacto de manera significativa en los estadísticos reportados.

1. **Tamaño base de datos**

La base final para la región 43 contiene 23.821 observaciones en total, con 14.481 del año 2005 y 9.340 del año 2025. No hay valores faltantes en la variable “pobre”. Se registran 2.900 hogares pobres (1.780 en 2005 y 1.120 en 2025) y 20.921 hogares no pobres (12.701 en 2005 y 8.220 en 2025). La cantidad de variables limpias y homogeneizadas es de 261, manteniendo consistencia en ambos períodos. Esto permite realizar análisis comparativos confiables entre los dos trimestres.

|  |  |  |  |
| --- | --- | --- | --- |
| **Concepto** | **2005** | **2025** | **Total** |
| Cantidad observaciones | 14.481 | 9.340 | 23.821 |
| Cantidad de observaciones con NAs en la variable "pobre" | 0 | 0 | 0 |
| Cantidad de Pobres | 1.780 | 1.120 | 2.900 |
| Cantidad de No Pobres | 12.701 | 8.220 | 20.921 |
| Cantidad de variables limpias y homogeneizadas | 261 | 261 | 261 |

Parte II: Métodos No Supervisados

PCA

Aplicamos PCA sobre las seis variables continuas (edad, edad2, educ, itf\_ajustado, ix\_tot, horastrab). Previamente estandarizamos (z-score) para que la escala no sesgue los componentes, y trabajamos con la matriz de correlaciones (no de covarianzas) tal como vimos en clase. Recordatorio: los componentes son ortogonales, el signo de las cargas es arbitrario y PCA no se interpreta como causalidad sino como reducción de dimensionalidad y ejes sintéticos de variación.

1

Tabla

El contenido generado por IA puede ser incorrecto. La matriz de correlaciones muestra una relación positiva entre el ingreso total familiar (ITF) y los años de educación (educ), sugiriendo que mayores niveles educativos se asocian con mayores ingresos. La edad presenta una correlación positiva moderada con el ingreso, aunque al considerar su cuadrado (edad²) la relación tiende a estabilizarse. Las horas trabajadas semanales y el tamaño del hogar muestran correlaciones bajas con el ingreso, lo que indica que el nivel de ingresos no depende linealmente de estas variables. En general, las correlaciones entre predictores son moderadas, lo que justifica el uso de PCA para reducir dimensionalidad sin pérdida significativa de información.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

2. El gráfico de dispersión entre el primer y segundo componente principal evidencia cierta separación entre los hogares pobres y no pobres. Los primeros componentes capturan la mayor parte de la varianza total, donde el primer componente parece estar fuertemente asociado con el ingreso y la educación, mientras que el segundo componente recoge variaciones relacionadas con la edad y el tamaño del hogar. Se observa que los hogares con menores ingresos tienden a concentrarse en valores bajos del primer componente.

Gráfico, Gráfico de líneas

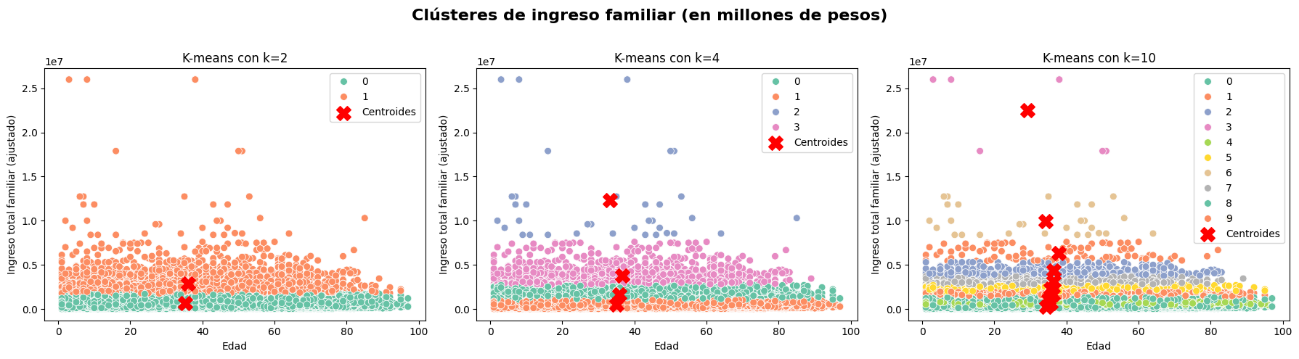
El contenido generado por IA puede ser incorrecto.

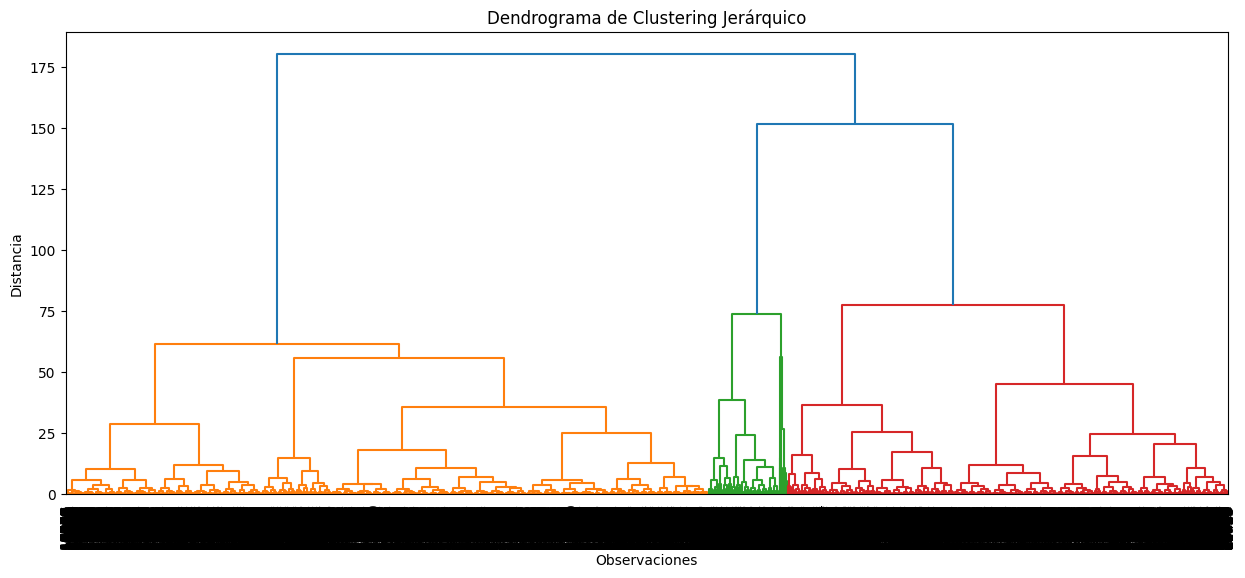
3.El gráfico de “loadings” muestra que las variables ingreso familiar y educación tienen los mayores pesos positivos en el primer componente, indicando que este eje refleja un gradiente de bienestar económico-educativo. Por su parte, las variables edad y edad² aportan de forma relevante al segundo componente, representando diferencias demográficas. Las variables horastrab y miembros del hogar tienen menor peso en la varianza total explicada.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

4 El primer componente explica alrededor del 40% de la varianza y el segundo el **23%**. Por eso, en conjunto, ambos concentran cerca del **64%** de la información. A partir del tercer componente el aporte adicional sería reducido, por lo que **trabajar con dos componentes** resulta suficiente.

1. Cluster
2. Cluster k-medias:
   1. Al realizar el análisis de clúster con dos centroides no se logra una identificación completa de las personas que caen bajo la línea de la pobreza. El resultado más limpio en cuanto a la clasificación de pobres-no pobres parecería ser el que se obtiene con cuatro clústeres ya que segmenta mejor el área de los ingresos mas bajos; por otro lado, en la iteración con k = 10 se logra una separación mas fina de los valores extremos lo que resulta valioso en el análisis, aunque incorporando varianza intracluster; lo anterior se muestra en el grafico a continuación.
   2. La exploración visual del número óptimo de clústeres a través del llamado método elbow (de codo) refuerza la conclusión de que el número optimo en la clasificación de grupos de ingreso se encuentra en 4 ya que a partir de este punto la disminución de la inercia se suaviza obteniéndose ganancias progresivamente menores, a pesar de que esta partición de la muestra da una visión bastante precisa de la separación de clases sociales por ingreso sobre todo en lo que concierne a la línea de pobreza definida, agregar más clústeres mejora la identificación de valores extremos, debe sopesarse la importancia de esta ultimo factor en el análisis contrastándolo con la pérdida de claridad.
3. Clúster jerárquico

Un dendograma es una representación atractiva de las observaciones que toma forma de árbol, comenzando por el nivel individual y agregando grupos cada vez más amplios de acuerdo a sus caracterisricas comunes, cada bifurcación representa la unión de dos clústeres y la altura indica la distancia o disimilitud entre ellos. Cortando el dendrograma a cierta altura, se puede obtener un número deseado de clústeres.

El clúster jerárquico por ingreso familiar y edad arroja tres grandes clasificaciones entre personas que se asemejan tanto en edad como en nivel de ingreso; a diferencia del algoritmo de k-medias a partir de este análisis no es posible distinguir a personas que se ubican por debajo por la línea de la pobreza como se muestra en la figura anterior.

1. Cluster k-Modas

Al tomar todas las variables categóricas para realizar la agrupación en pobres y no-pobres con el número de clústeres predefinidos para el algoritmo de k-modas podemos observar que con k=2 se logra aproximar de manera bastante precisa la proporción de pobres encontrada en el TP1 de cerca de 12% aunque no con el mismo número de observaciones; al aumentar el número de clústeres los resultados se progresivamente mas variables.

