UNIVERSIDAD DE BUENOS AIRES - 2025

Taller de Programación

TRABAJO PRÁCTICO N° 2

HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

Fecha de entrega: 22 de octubre a las 13:00 hs.

Contenido: Continuar con la familiarización con la base de datos de la Encuesta Permanente de Hogares. Hacer una ejercitación de Histogramas & Kernels y los métodos no supervisados vistos en clase (PCA & Cluster).

Modalidad de entrega

- Asegurense de haber creado una carpeta llamada TP2 en el repositorio de GitHub de cada grupo.

- El informe debe subirse a dicha carpeta en repositorio del grupo en formato PDF con el nombre **Program_TP2_Grupo#.pdf** (donde **#** es el número de grupo), incluyendo gráficos e imágenes dentro del mismo archivo. La extensión máxima es de 5 páginas (sin apéndices) y se espera una redacción clara y precisa.

- Se debe publicar el código con los comandos utilizados, indicando claramente a qué inciso corresponde cada uno. El nombre del archivo deberá ser **Program_TP2_Grupo#**.
    - Al finalizar el trabajo práctico deben hacer un último commit en su repositorio de GitHub llamado "Entrega final del tp".
    - El Jupyter Notebook y el correspondiente al TP2 deben estar dentro de esa carpeta.
    - La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
        - No suban el pdf en la sección de "**Actividades/Entregas**" del campus hasta no haber terminado y estar seguros de que han hecho el *commit* y *push* a la versión final que quieren entregar.
        - No hagan nuevos *push* después de haber entregado su versión final. Esto generaría confusión acerca de qué versión es la que quieren que se les corrija.

- Deben subir el trabajo práctico en en la sección de "**Actividades/Entregas**" con el título de entrega "**TP 2 - Grupo #"**.

- Además, deben adjuntar en el documento del informe el link del repositorio público del Github donde tienen los códigos que utilizaron para resolver el trabajo práctico.

- En resumen, la carpeta del repositorio debe incluir:

    - El código

    - Un documento PDF donde están las figuras y una breve descripción de las mismas.

- **Cualquier detección de copia o plagio será sancionada.**

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

La idea de esta primera parte es que continuen con la limpieza de la base de datos que contiene las observaciones del primer trimestre de 2005 y del primer trimestre de 2025. La **base final** a trabajar resultante debe incluir todas las variables en ambos trimestres que limpiaron en el TP anterior y las nuevas que se les pide en este TP, expresadas de manera homogénea. Es decir, si la variable CH04 en 2005 toma los valores "Hombre" o "Mujer", y en 2025 toma los valores 1 y 2, la variable limpia en la **base final** debe tener solamente dos valores consistentes.

1) Cree la variable "*edad2*" definida como $edad^2$ (edad al cuadrado). Presente un histograma de la variable edad en un panel A, y a la par una distribución de kernels para los pobres y no pobres en un panel B (esto es, son dos líneas de kernel en este segundo panel). Comente brevemente la distribución de edades en estos dos paneles (3-4 oraciones).

2) Cree la variable *educ* definida como la cantidad de años de educación. Use inteligentemente las variables CH12, CH13 y CH14 para crearla. Por ejemplo, si dice que el nivel más alto de educación es "Secundario" (CH12), "Sí" finalizo este nivel (CH13) y el último año que aprobó (CH14) fue "sexto", entonces puede asumir que tiene *educ=12*, osea 12 años de educación formal. Presente una estadística descriptiva (promedio, sd, min, p50, max) de dicha variable creada y comente

3) Actualice la variable *ingreso_total_familiar* con el total de ingresos habituales (ITF). Recuerde que los pesos de 2005 tienen un poder de compra distinto a los pesos de 2025 en el primer trimestre. Convierta primero los ingresos de 2005 a pesos de 2025. Similar al ítem 1, presente en un panel A, un histograma de la variable *ingreso_total_familiar* y las distribuciones de kernels para pobres y no pobres en un panel B. Comente brevemente la distribución de ingresos en estos dos panels (3-4 oraciones). En cada panel, sume una linea vertical con la línea de la pobreza calculada en el TP1.

4) Para el jefe del hogar, cree la variable *horastrab* como el total de horas trabajadas como la suma de las horas en la ocupación principal y otras ocupaciones (PP3E_TOT + PP3F_TOT). Presente una estadística descriptiva (promedio, sd, min, p50, max) de dicha variable creada y comente

5) ¿Cuál es el tamaño de la de la base de datos **para su región** con las variables originales unificadas? Para ello complete la tabla 1 que se le diseña abajo y comente.

Tabla 1. Resumen de la base final para la región YYY

	2005	2025	Total
Cantidad observaciones			
Cantidad de observaciones con NAs en la variable "Pobre"			
Cantidad de Pobres			
Cantidad de No Pobres			
Cantidad de variables limpias y homogeneizadas			

Nota: Calcular la cantidad de pobres y no pobres a partir de la variable de *Pobre* que crearon en el trabajo práctico 2.

Parte II: Métodos No Supervisados

Esta parte del trabajo práctico tiene como objetivo que realicen un análisis visual de los datos utilizando las herramientas vistas en clase. En esta parte, solo necesita utilizar las variables: edad, edad2, educ, ingreso_total_familiar (ITF), el número de miembros en el hogar (2005=IX_TOT y 2025=IX_Tot) y horastrab.

1. Realice una matriz de correlaciones con estos seis predictores para su región y comente los resultados.

A. PCA

2. PCA con ingreso: Apliquen PCA a las seis variables seleccionadas para esta parte. Recuerde primero estandarizar las variables como vimos en la tutorial. En un gráfico de dispersión muestren los índices (*scores*) calculados del primer y segundo componente de PCA y comente los resultados.

3. Grafique con flechas los ponderadores (*loading*) de PCA para el primer y segundo componente y comente los pesos que que le dan a cada variable utilizada.

4. Finalmente, grafique la proporción de la varianza explicada para cada uno de los seis componentes y comente el gráfico.

B. Cluster

5. Cluster k-medias:
    a. Corran el algoritmo con $k = 2$, $k = 4$ y $k = 10$ usando n_init = 20, y grafiquen los resultados usando edad e ingreso familiar.

Interprétenlos ¿Puede el algoritmo con $k = 2$ separar correctamente a las personas pobres y no pobres en su región?

b. Grafique alguna medida de disimilitud para $k = 1$ hasta $k = 10$. Usando la inspección visual de *Elbow* ¿cuál sería el número óptimo de clusters en su región? ¿Dicha cantidad de grupos nos ayudaría a distinguir entre *pobres* y *no pobres* o entre distintas clases socioeconómicas?

6. Cluster jerárquico: Utilizando las variables mencionadas arriba, realicen un análisis de clustering jerárquico. Generen un dendograma y expliquen brevemente qué es un dendograma.

7. (Opcional) Cluster k-moda: Implemente cluster k-moda con $k = 2$, $k = 4$ y $k = 10$ y todas las variables **dummies** de la EPH que tiene en su base de datos limpias (excepto la categorica pobre y no pobre). ¿Puede el algoritmo con $k = 2$ separar correctamente a las personas pobres y no pobres en su región?