UNIVERSIDAD DE BUENOS AIRES - 2025

BIG DATA & MACHINE LEARNING

TRABAJO PRÁCTICO N° 3

HISTOGRAMAS, KERNELS & MÉTODOS NO SUPERVISADOS USANDO LA EPH

Fecha de entrega: Mayo 13 de marzo a las 13:00 hs.

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

Modalidad de entrega

- Asegurense de haber creado una carpeta llamada TP3 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 **Big_Data_TP3_Grupo#.pdf** (donde **#** es el número de grupo), incluyendo gráficos e imágenes dentro del mismo archivo. La extensión máxima es de 8 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 **Big_Data_TP3_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 TP3 deben estar dentro de esa carpeta.
    - La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
        - No envien el correo 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* despues de haber entregado su versión final. Esto generaría confusión acerca de que versión es la que quieren que se les corrija.

- También deben enviar el link de su repositorio -para que pueda ser clonado y corregido- a mi correo 25RO35480961@campus.economicas.uba.ar . Usar de asunto de email "**Big Data - TP 3 - Grupo #**" donde # es el número de grupo que le fue asignado.

- En resumen, la carpeta del repositorio debe incluir:

    - El codigo

    - Un documento Word donde esten las figuras, tablas y una breve descripción de las mismas. Tenga en cuenta la devolución general de TP2. Si no incorpora dichos lineamientos el nuevo reporte, entonces se le descontara puntos por mala presentación de la información.

- **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 completen la limpieza de la base de datos que contiene las observaciones del primer trimestre de 2004 y del primer trimestre de 2024. La **base final** a trabajar resultante debe incluir todas las variables presentes en ambos trimestres, expresadas de manera homogénea. Es decir, si la variable CH04 en 2004 toma los valores "Hombre" o "Mujer", y en 2024 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 ocupados y desocupados en un panel B (esto es, son dos lineas de kernel en este segundo panel). Comente brevemente la distribución de edades en estos dos panels (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 ultimo año que aprobó (CH14) fue "sexto", entonces puede asumir que tiene *educ=12*, osea 12 años de educación formal. Presente una estadistica descriptiva (promedio, sd, min, p50, max) de dicha variable creada y comente

3) Cree la variable *salario_semanal* como el total de ingresos habituales (P21) dividido 40. Esta variable nos da una aproximación del salario semanal suponiendo que la persona trabaja a tiempo completo 8 horas al dia, 5 dias a la semana (8x5=40). Sin embargo, antes de hacer dicha división recuerde su bonus de *economista*. Los pesos de 2004 tienen un poder de compra distinto a los pesos de 2024 primer trimestre. Convierta primero los ingresos de 2004 a pesos de 2024.

    a) Similar al item 1, presente en un panel A, un histograma de la variable *salario_semanal* y las distribuciones de kernels para ocupados y desocupados en un panel B. Comente brevemente la distribución de salarios en estos dos panels (3-4 oraciones).

4) 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 estadistica 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 region YYY

	2004	2024	Total
Cantidad observaciones			
Cantidad de observaciones con Nas en la variable "Estado"			
Cantidad de Ocupados			
Cantidad de Desocupados			
Cantidad de variables limpias y homogeneizadas			

Nota: Se calcula la "cantidad de Ocupados" como aquellos con la variable "Estado==Ocupado" y Cantidad de Desocupados como aquellos con la variable "Estado==Desocupado".

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, salario_semanal y horastrab.

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

A. PCA

2. PCA con salario: Apliquen PCA a las cinco variables seleccionadas para esta parte. Recuerde primero estandarizar las variables como el la Clase 6. En un gráfico de dispersión muestre los indices (*scores*) calculados del primer y segundo componente de PCA y comente los resultados.

3. Grafique con flechas los pronderadores (*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 para cada uno de los componentes y comente el grafico.

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 dos predictores. Interprétenlos.

b. Grafique *edad* y *educ* de los resultados de $k = 2$ donde cada punto tome dos colores, un color para *ocupados* y otro color para *desocupados*. ¿Puede el algoritmo separar correctamente a las personas ocupadas de las desocupadas? Comenten.

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