**Big Data y Machine Learning (UBA) - 2025**

**Trabajo Práctico 3: Histogramas, Kernels & Métodos No Supervisados usando la EPH**

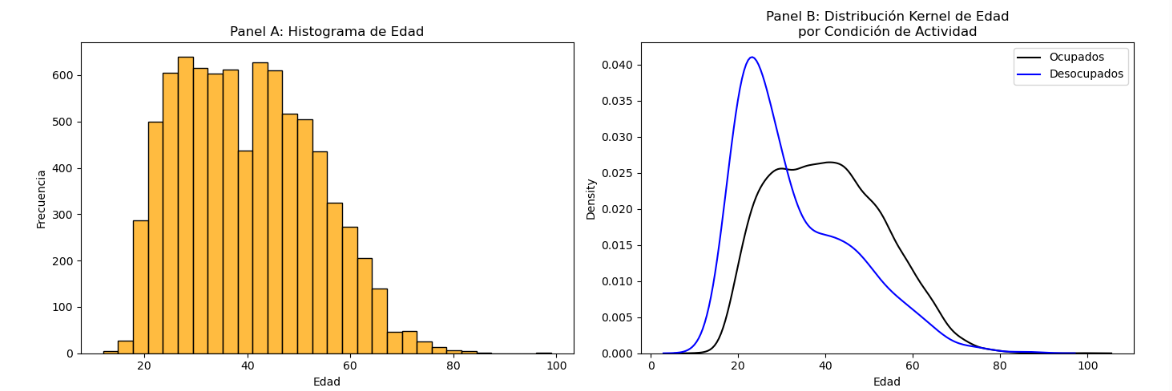
**Materia: Big Data y Machine Learning**

**Profesora: Romero Noelia**

**Grupo:18**

**Integrantes:**

|  |  |  |
| --- | --- | --- |
| 911748 | 44486233 | MAMONE FACUNDO |
| 915896 | 46278154 | ROMERO MARTIN |
| 916370 | 96256981 | VARGAS VILLA AROON RICARDO |

**Al igual que en el TP N2 se trabajó con la región Patagonia. Parte I-**

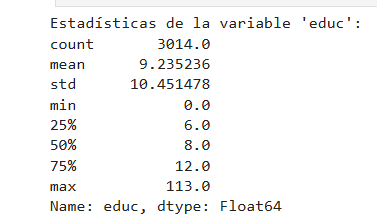
**Gráfico A**-El histograma representa la distribución de frecuencias de la variable edad dentro de la muestra analizada se observa que la distribución presenta una forma asimétrica negativa mayor concentración de individuos en el rango **entre 25 y 50 años.** A partir de los 50 años, la frecuencia comienza a disminuir progresivamente: alta representación de adultos jóvenes y una menor presencia de personas en edades avanzadas. El número de observaciones cae significativamente después de los 60 años, lo que puede reflejar procesos de jubilación o retiro del mercado laboral.

**Gráfico B**- Esta muestra la densidad de probabilidad diferenciada por condición de actividad: **ocupados (línea negra)** y **desocupados (línea azul).**

La curva de los ocupados presenta un pico amplio y relativamente plano entre los 30 y 50 años, indicando que una gran parte de la población ocupada se encuentra en ese rango. Esto es coherente con el ciclo laboral típico, en el cual los individuos logran estabilidad y consolidación en el mercado de trabajo durante esos años.

La distribución de los desocupados es marcadamente distinta: su pico se encuentra alrededor de los 20 años, y decrece rápidamente a medida que aumenta la edad Esto indica que la desocupación afecta desproporcionadamente a los jóvenes dándose por distintos factores.

1. **Tabla 1**

En trabajada tabla podemos observar que la variable “eudc” representa la cantidad estimada de años de educación formal elaborada a partir de las variables solicitadas.

-Aquí visualizamos que la media es de 9,24 años, mientras que la mediana es de 8 años, esta diferencia indica una asimetría positiva, es decir, hay una proporción pequeña de individuos con niveles educativos muy altos que elevan el promedio. El desvío estándar es alto (10,45), lo cual sugiere una gran heterogeneidad educativa dentro de la muestra. (El valor máximo de 113 años es anómalo y probablemente se trate de un error de carga de datos). La presencia de valores nulos o cero (mínimo = 0) indica que hay personas sin ningún año formal de educación. En conclusión, El nivel educativo promedio está por debajo del nivel secundario completo, lo que puede impactar negativamente en la inserción laboral y los niveles de ingreso.

3)

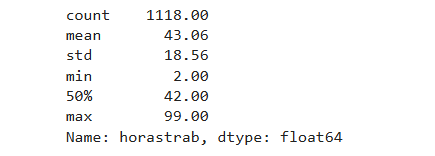
Gráfico, Histograma

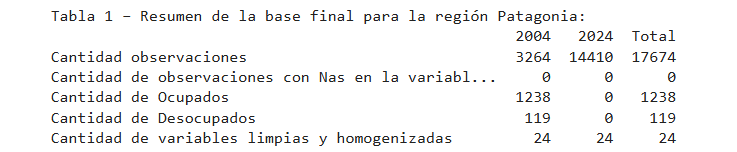
El contenido generado por IA puede ser incorrecto.

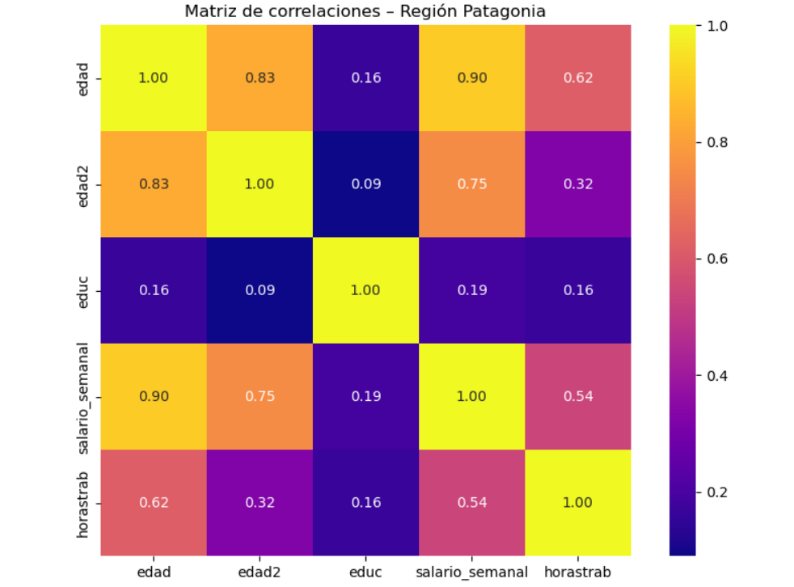
**Grafico A**- Este gráfico muestra la distribución de la variable salario semanal ajustado, expresada en pesos constantes de 2024. Para ello utilizamos una calculadora de inflación y establecimos que $1 del 2004 equivale a $1000 del 2024. El mayor número de observaciones (casi 400) se encuentra en el primer tramo (salarios bajos), lo que sugiere que una parte significativa de la muestra recibe ganancias bajas. A medida que el salario aumenta, la frecuencia decrece rápidamente, la cantidad de personas en esos rangos cae drásticamente, lo que indica que los altos salarios son excepcionales dentro de la muestra.

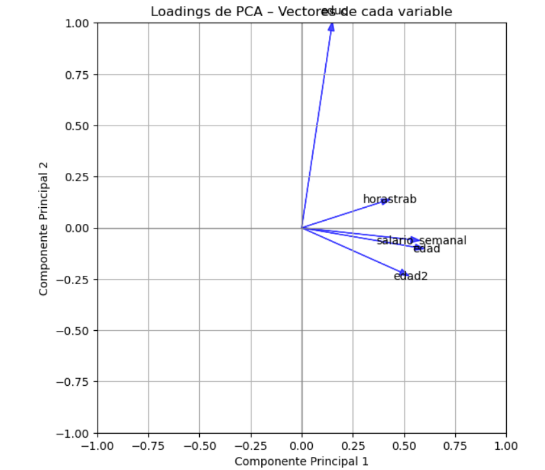
**Grafico B**- La curva correspondiente a los ocupados tiene un pico alrededor de los 15.000 a 20.000 pesos semanales, lo cual refleja una distribución más dispersa y una mayor presencia relativa de ingresos elevados en este grupo.

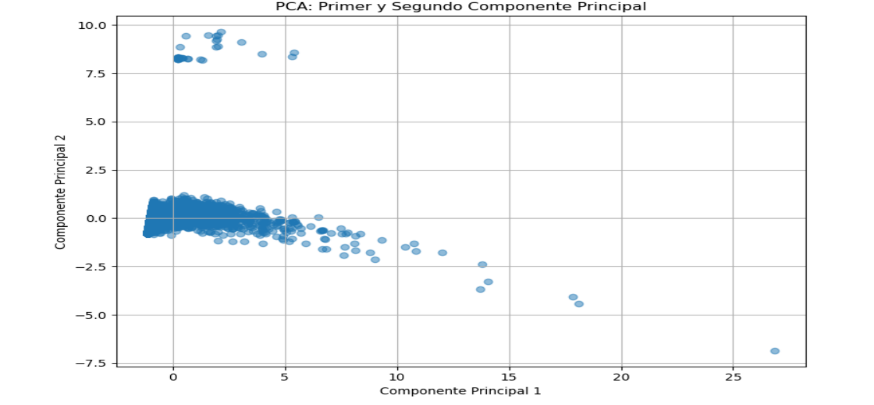
la curva de los desocupados está altamente concentrada en valores cercanos a cero, con una caída rápida de densidad a medida que el salario aumenta. Esto es coherente, ya que por definición los desocupados no perciben ingresos laborales.

**4) Tabla 2** La media (43,06 horas) y la mediana (42 horas) están muy próximas, lo cual sugiere una distribución bastante igualitaria en torno al valor típico de una jornada laboral completa aunque si nos centramos en el desvío estándar alto (18,56 horas) esta revela una alta variabilidad en la cantidad de horas trabajadas. Esto puede deberse a la coexistencia de trabajadores a tiempo parcial, informalidad, pluriempleo o sobrecarga laboral quedando en evidencia la gran El desvío estándar alto (18,56 horas) revela alta variabilidad en la cantidad de horas trabajadas. Esto puede deberse a la coexistencia de trabajadores a tiempo parcial, informalidad, pluriempleo o sobrecarga laboral. Haciendo foco en el máximo de 99 horas, un caso totalmente atipico, podría darse lo que ya mencionamos anteriormente, indicios de explotación o informalidad extrema.

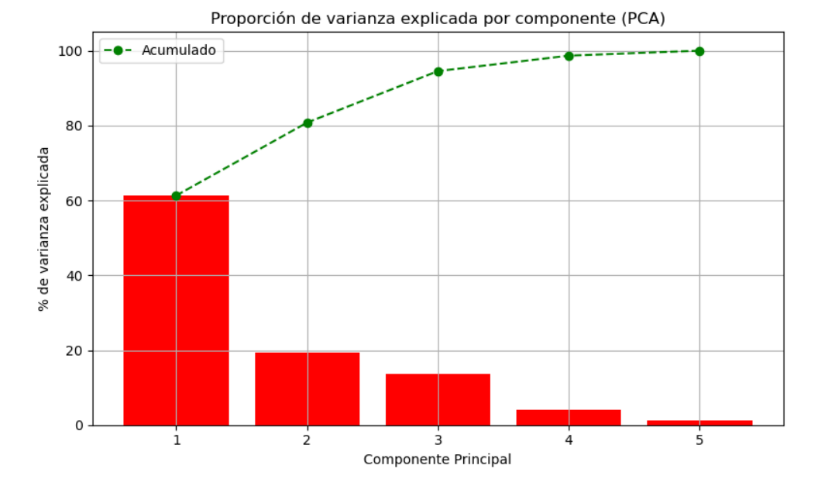
**5) Tabla 3-** La muestra total incluye 17.674 observaciones, con un total de 1238 ocupados para el año 2004 el cual representa un gran porcentaje de personas en actividad laboral comparado a las 119 desocupadas. El total de personas activas (ocupadas + desocupadas) en 2004 asciende a 1.357, lo que representa aproximadamente el 41,5% de las observaciones para ese año. Esto sugiere que el resto de la población puede estar conformado por inactivos (estudiantes, jubilados, personas dedicadas a tareas del hogar, etc.). Además podemos observar que no hay valores faltantes (NaN) en las variables consideradas, lo que garantiza la consistencia y robustez estadística de los análisis posteriores.

**Parte II**-6)

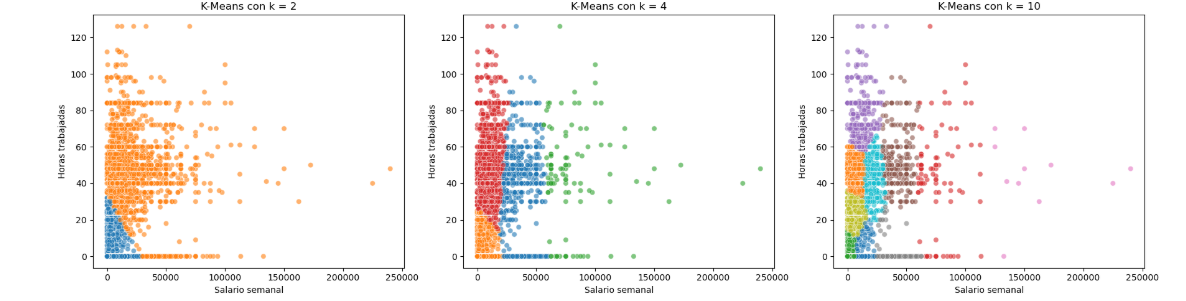
En esta matriz trabajada vemos que Edad y salario\_semanal tienen una correlación fuerte (0.90), lo cual es totalmente lógico ya que muchas veces a mayor edad, mayor salario, lo que puede reflejar la acumulación de experiencia laboral. También observamos que Edad y edad2 tienen una correlación muy alta (0.83), lo que es esperable al tratarse de una relación funcional. Por otro lado, la educación tiene correlaciones bajas con todas las variables (máximo 0.19), lo que sugiere que en esta región y año, el nivel educativo no está fuertemente correlacionado con el salario o las horas trabajadas, ya que lo mencionamos, salario y horas trabajadas tienen una correlación moderada (0.54), reflejando que trabajar más horas puede asociarse con mayores ingresos, pero no de forma lineal ni proporcional.

7-8)

Trabajado gráfico, PCA, lo que busca hacer es reducir la dimensionalidad del conjunto de variables conservando la mayor parte de la varianza. Podemos observar el Componente Principal 1 (PC1) parece captar la mayor variabilidad entre los individuos, fuertemente asociado a “salario\_semanal”, “edad” y “horastrab”, lo que confirma que este componente explica la varianza asociada a la intensidad y nivel de inserción laboral, mientras que, el Componente 2 (PC2) permite distinguir un subconjunto de individuos que se separan fuertemente en la parte superior del gráfico, observando que en el eje vertical (PC2) está dominado casi exclusivamente por “educ”. El grafico por ejemplo, muestra que la edad y el salario están fuertemente correlacionados en la región, entre otras cosas.

9)

Este gráfico muestra un análisis de varianza explicada por cada componente principal del PCA. En rojo, las barras indican la varianza explicada individual por cada componente, y la línea verde punteada indica el acumulado. Podemos observar que el PCA es altamente efectivo ya que con los primeros 2 componentes ya se conserva más del 80% de la varianza lo que justifica su uso para reducción de dimensionalidad sin pérdida significativa de información, a partir de allí, el aporte marginal es cada vez menor: PC3 (~13%), PC4 (~5%) y PC5 (menos del 1%) ya no suman mucho valor adicional al análisis.

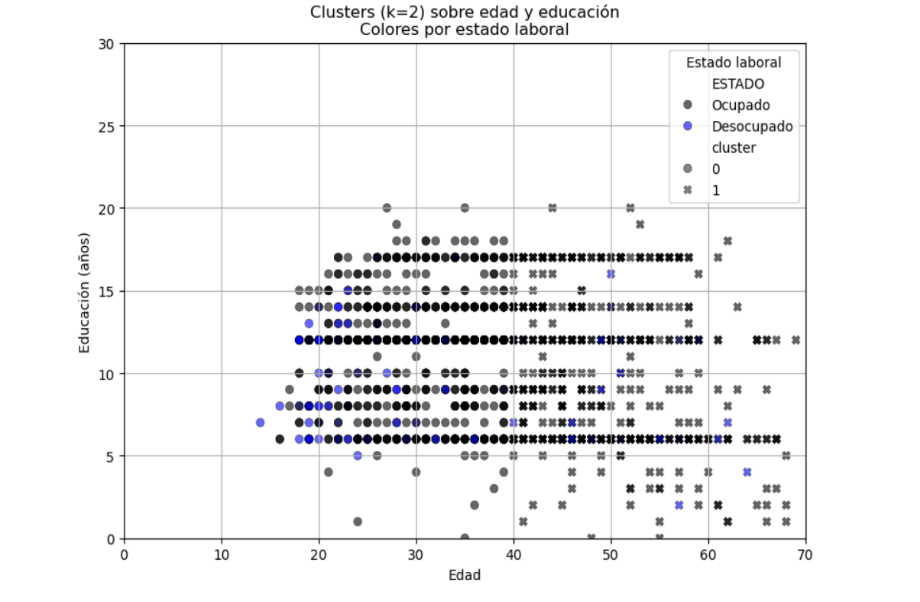
10) 

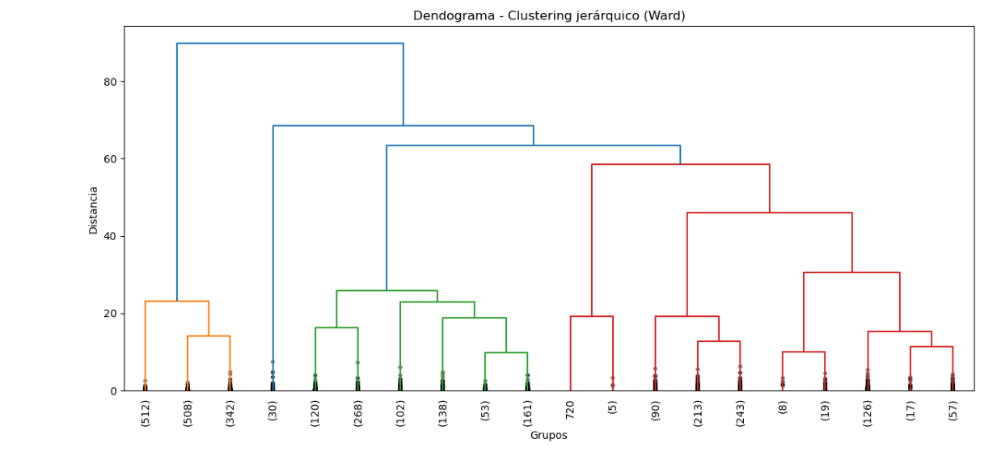
Aquí podemos observar que con k=2 se identifican dos grandes grupos, un grupo de trabajadores de bajos ingresos, con una amplia dispersión en horas trabajadas y Otro grupo, más reducido, que concentra los trabajadores con salarios relativamente más altos.

En el panel central k=4, el algoritmo detecta mayor heterogeneidad interna separando, personas de bajos ingresos y pocas horas trabajadas, con altas horas trabajadas y bajos ingresos, con ingresos intermedios y estables y un grupo reducido de ingresos muy altos, que representa una minoría privilegiada. Algunas de estas condiciones se pueden dar por posibles casos de subocupación o informalidad, entre otros.

Por ultimo k=10, es útil para identificar objetivos específicos o patrones atípicos, pero puede sobreajustar si no se justifica con criterios teóricos o métricas correspondientes.

11)

 Este gráfico representa los resultados de un clustering con k=2 sobre dos variables, edad (en eje X) y educación en años (en eje Y). Los puntos están coloreados según estado laboral, ocupado (negro) y desocupado (azul), y marcados por cluster (gris vs. negro). Podemos observar que hay 2 grandes grupos, mayoritariamente ocupados, con una amplia dispersión en edad (25 a 60 años) y nivel educativo (6 a 18 años). Refleja un perfil consolidado laboralmente, de adultos con escolaridad media o alta y por otro lado concentrado en jóvenes (menores de 30 años) con niveles educativos bajos (menos de 12 años). Presenta mayor proporción de desocupados, lo que refuerza la vulnerabilidad laboral de jóvenes con baja formación lo cual en conjunto da a entender que la educación y la edad son claves en la segmentación del empleo.

12)

Este grafico se utilizo la estructura jerárquica resultante del algoritmo de clustering jerárquico con método de Ward utilizando distintas variables para su posterior análisis, se observan tres macrogrupos principales (marcados por ramas azules, verdes y rojas), dentro de cada macrogrupo se identifican subgrupos más pequeños, de esta forma permitiendo explorar la estructura interna de los datos, para poder concluir en que existen estructuras naturales de agrupamiento en los datos, con subpoblaciones bien diferenciadas.