Big Data & Machine Learning - Trabajo Práctico N°3

Grupo: 1
Integrantes:

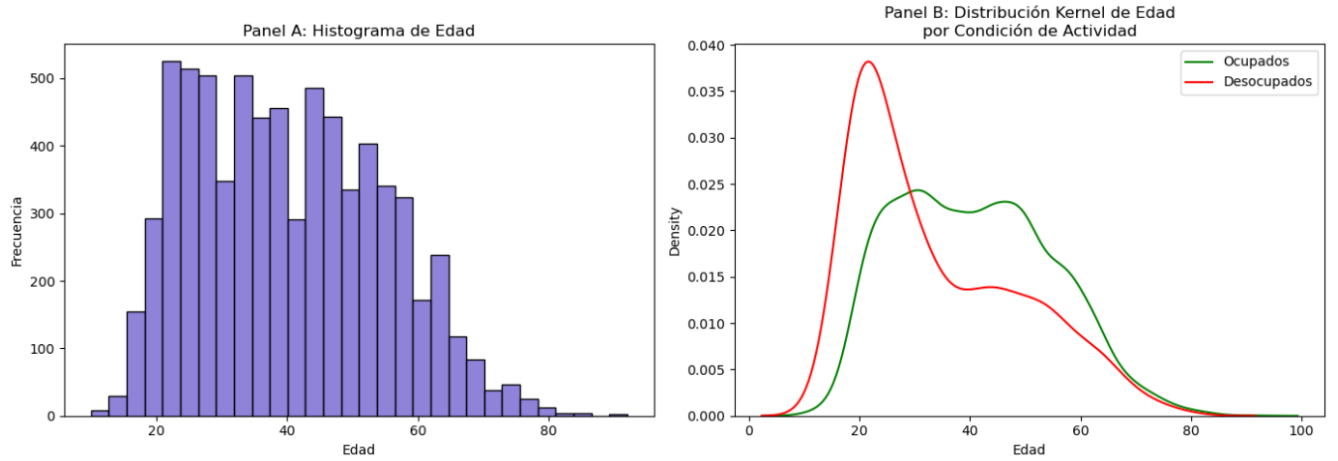
-Benjamin Rodolfo Ayay Quispe (N° Registro 912662)

-Jesús David Ochochoque Mendoza (N° Registro 915690)

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

1. Creación de variable “edad2”, histograma y kernel

Figura 1 y 2 - Región seleccionada: Gran Buenos Aires



En la figura 1 se muestra que la distribución de edades se concentre entre los 20 y 55 años, con una mayor frecuencia en los tramos 20-30 y 30-40. A partir de los 60, la participación disminuye marcadamente, lo que es esperable dado que muchas personas se retiran del mercado laboral. También se observa un leve sesgo a la derecha, con pocos casos por encima de los 70 años. La distribución global refleja la estructura demográfica típica de una población económicamente activa.
En la figura 2 las curvas de densidad muestran diferencias claras entre ocupados y desocupados. La distribución de los ocupados se concentra principalmente entre los 30 y 50 años, lo que indica que la ocupación es más frecuente en edades laborales intermedias. Por otro lado, la distribución de los desocupados tiene un pico más pronunciado entre los 18 y 30 años, señalando que la desocupación afecta más a los jóvenes. Esto sugiere una posible barrera de entrada al mercado laboral para los recién incorporados, o inestabilidad en sus empleos.

2. Creación de variable “educ”, estadística descriptiva

Tabla 1

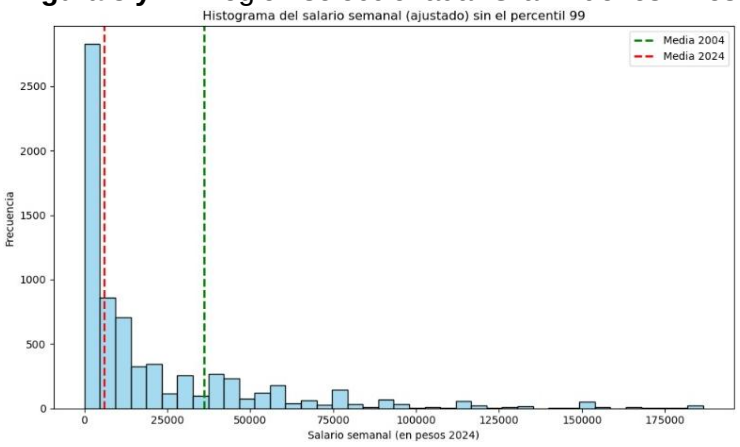
ESTADÍSTICO	VALOR
Cantidad	10,418.00
Media	7.75
Desvío estándar	5.47
Mínimo	0.00
Mediana (p50)	7.00
Máximo	23.00

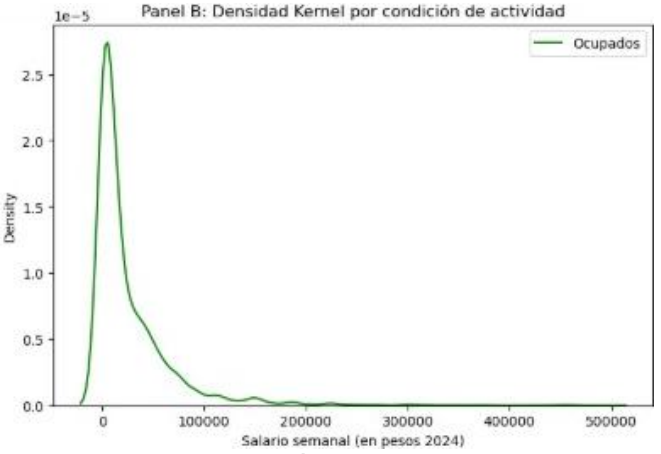
En la tabla 1, la variable educ representa la cantidad estimada de años de educación formal completados por cada individuo, construida a partir de las variables CH12, CH13 y CH14. La distribución de esta variable muestra una media de 7,75 años, lo que indica que en promedio la población no alcanza a completar el nivel secundario completo. La mediana es de 7 años, lo que refuerza esta observación. El mínimo es 0, mientras que el máximo registrado es 23 años, lo cual puede representar trayectorias educativas prolongadas como posgrados. El desvío estándar de 5,47 refleja una dispersión moderada, consistente con la heterogeneidad educativa del mercado laboral argentino.

3. Creación de variable “salario_semanal”

a. Histograma y kernel

Figura 3 y 4 - Región seleccionada: Gran Buenos Aires





En la figura 3 se representa el histograma del salario semanal de la población ocupada y desocupada, expresado en pesos constantes de 2024. Para mejorar la visualización y eliminar valores atípicos, se excluyeron las observaciones por encima del percentil 99 (el 1% con mayores ingresos). La distribución resultante muestra una marcada concentración en los tramos salariales más bajos, lo que evidencia una fuerte asimetría a la derecha, típica de las variables de ingreso. Las líneas verticales indican los valores promedio del salario semanal para los años 2004 y 2024. Se observa un desplazamiento hacia la derecha en la media de 2024, reflejando un aumento del ingreso medio real a lo largo del período analizado.
En la figura 4 se observa que los ocupados presentan una distribución que comienza cerca de cero y tiene un pico alrededor de los ingresos bajos, seguido por una caída progresiva. Dado que los desocupados no tienen ingresos laborales, su salario semanal es 0. En la densidad kernel, esto implica que no presentan variabilidad para estimar una curva, por lo que solo se visualiza la distribución de los ocupados.

4. Creación de variable “horastrab”, estadística descriptiva

Tabla 2

ESTADÍSTICO	VALOR
Cantidad	5,672.00
Media	39.55
Desvío estándar	18.57
Mínimo	1.00
Mediana (p50)	40.00
Máximo	100.00

En la tabla 2, la distribución de horas trabajadas presenta una media de 39.5 horas semanales con una mediana de 40, lo que sugiere una predominancia de jornadas completas cercanas al estándar laboral. Sin embargo, la desviación estándar revela una heterogeneidad significativa en el mercado laboral, evidenciando la coexistencia de trabajadores con empleos parciales hasta 1 hora y sobreocupados hasta 100 horas. El hecho de que el 50% de la muestra concentre exactamente 40 horas refleja la importancia de esta jornada, mientras que el máximo de 100 horas podría indicar fenómenos de pluriempleo o informalidad.

5. Tabla completa

Tabla 3

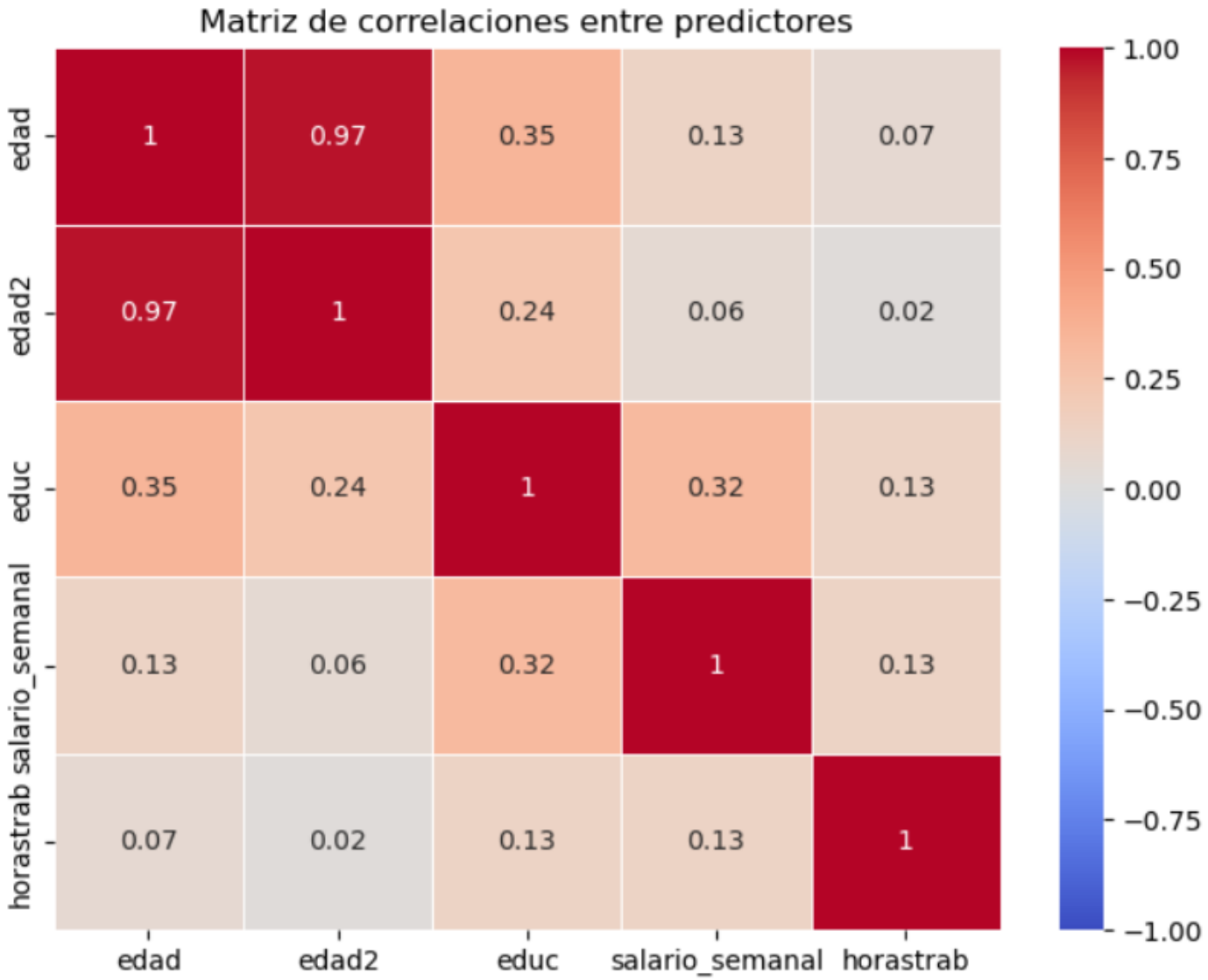
	2004	2024	TOTAL
Cantidad de observaciones	7647	7051	14698
Cantidad de observaciones con Nas en la variable “Estado”	0	0	0
Cantidad de ocupados	3079	3224	6303
Cantidad de desocupados	528	311	839
Cantidad de variables limpias y homogenizadas	13	13	26

En la tabla 3, la muestra consolidada contiene 14,698 observaciones, con 6,303 ocupados (42.9% del total) y 839 desocupados (5.7%), evidenciando una estructura laboral donde la población ocupada predomina significativamente. La ausencia de valores nulos en la variable ESTADO (0 NaN) garantiza confiabilidad en el análisis de participación laboral. Al comparar ambos años, se observa un incremento en la ocupación (2004: 40.3% vs 2024: 45.7%) y una reducción en la desocupación (2004: 6.9% vs 2024: 4.4%), lo que podría reflejar mejoras en las condiciones del mercado laboral en la región estudiada.

Parte II: Métodos No Supervisados

1. Matriz de correlaciones

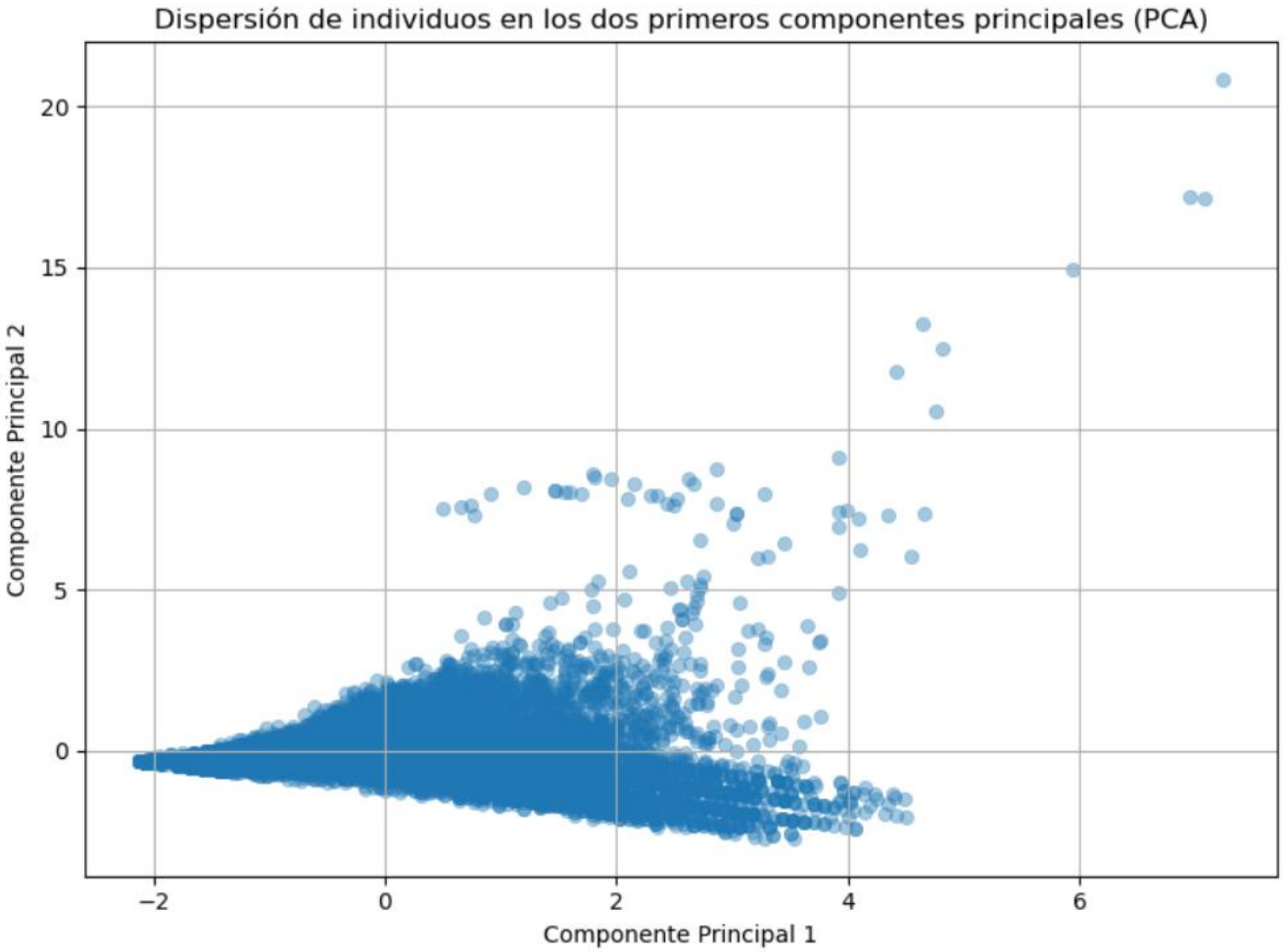
Figura 5 - Región seleccionada: Gran Buenos Aires



En la figura 5 se observa que existe una alta correlación positiva entre edad y edad2, como es lógico, ya que una es el cuadrado de la otra. La correlación entre educ y salario_semanal es positiva pero moderada, indicando que más educación tiende a asociarse con mayores ingresos, aunque no de forma perfecta. Además, horastrab y salario_semanal también presentan una correlación positiva, aunque puede verse afectada por variabilidad en el tipo de empleo (jornada parcial vs. completa).

2. Gráfico de dispersión

Figura 6 - Región seleccionada: Gran Buenos Aires

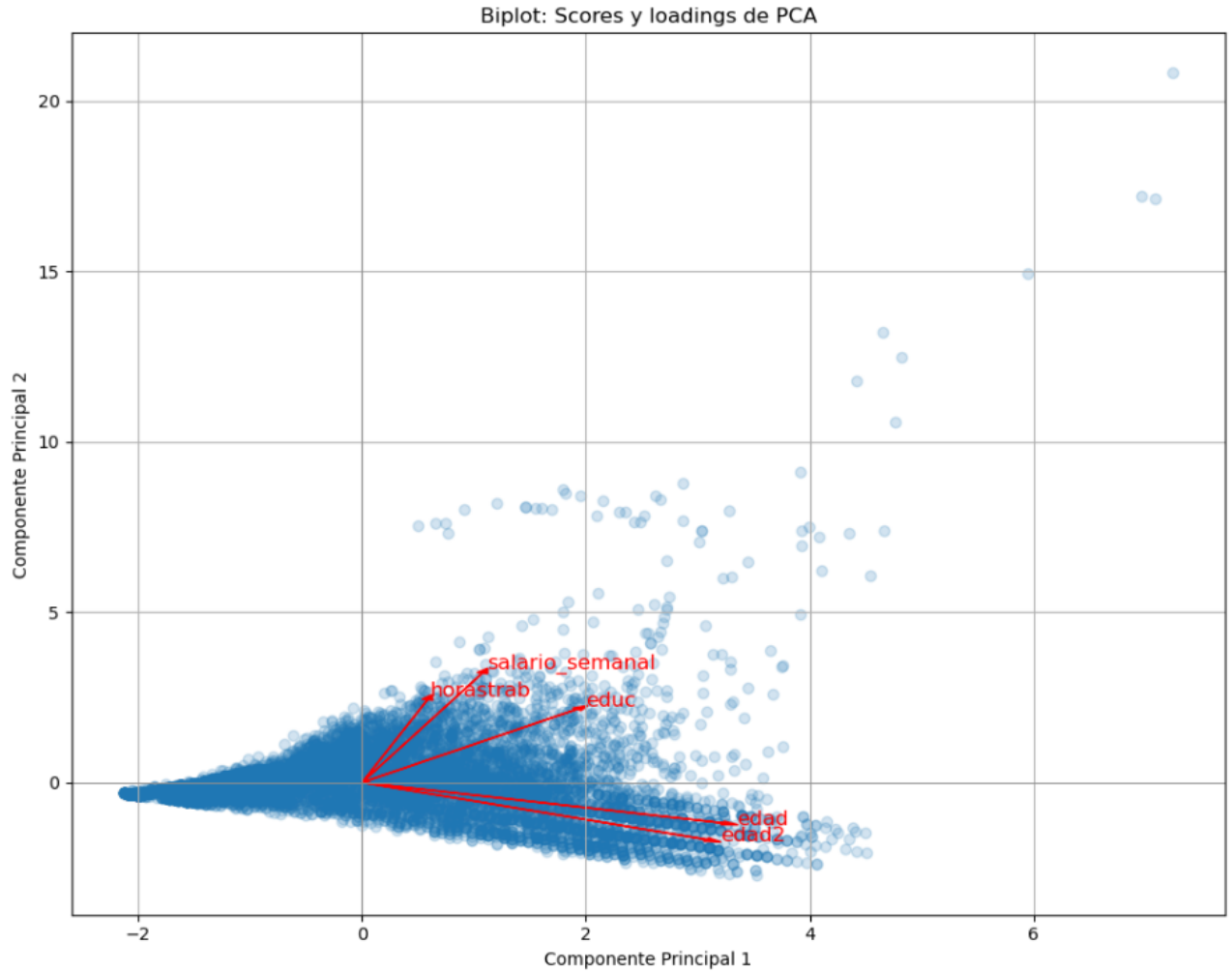


En la figura 6 aplicamos PCA para reducir las 5 variables correlacionadas a dos componentes principales ortogonales. El primer componente (PC1) captura la mayor parte de la variación total, relacionada probablemente con factores socioeconómicos como ingreso, educación y horas trabajadas. El segundo componente (PC2) captura otra dimensión independiente, como edad o intensidad laboral. El gráfico de dispersión muestra cómo se distribuyen los individuos según estas combinaciones, permitiendo detectar patrones latentes sin perder demasiada información.”

3. Gráfico de ponderadores

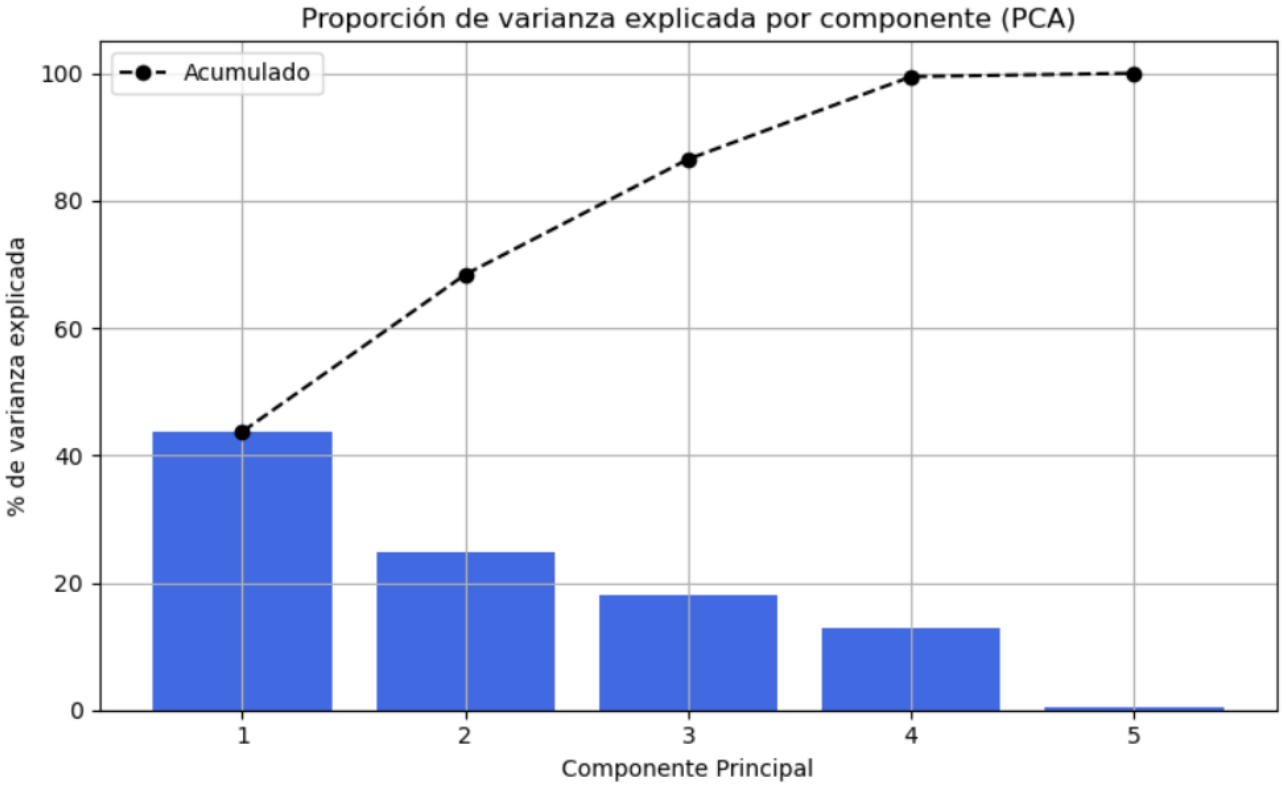
En la figura 7 se observa que los vectores del biplot revelan que el salario_semanal y educ tienen las mayores contribuciones al Componente Principal 1, con ponderadores positivos, confirmando que estas variables son los principales drivers de la variabilidad en los datos. Por otro lado, horastrab y edad muestran una influencia moderada en el Componente Principal 2, aunque con menor peso relativo. La variable edad2 sigue estrechamente a edad, como era esperable. La dirección similar de salario_semanal y educ sugiere una correlación positiva entre ambas, mientras que horastrab aparece casi ortogonal, indicando que su aporte a la varianza es independiente del salario/educación.

Figura 7 - Región seleccionada: Gran Buenos Aires



4. Gráfico de proporción de la varianza

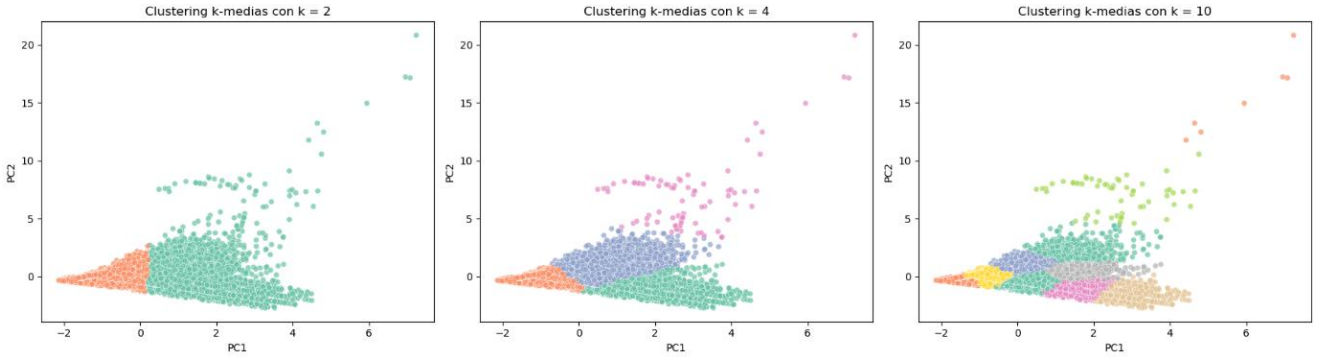
Figura 8 - Región seleccionada: Gran Buenos Aires



En la figura 8 se observa que el gráfico muestra que los dos primeros componentes principales explican aproximadamente el 68% de la varianza total del conjunto de datos. Esto indica que se puede reducir el número de dimensiones de 5 a 2 manteniendo gran parte de la información. El tercer componente agrega valor si se quiere capturar más detalle, llevando el total acumulado a más del 85%. A partir del cuarto componente, la ganancia marginal de información es baja, y el quinto componente es despreciable. Esto respalda el uso de 2 o 3 componentes para análisis visuales y clustering, sin pérdida significativa de contenido informativo.

5a. Gráfico usando dos predictores

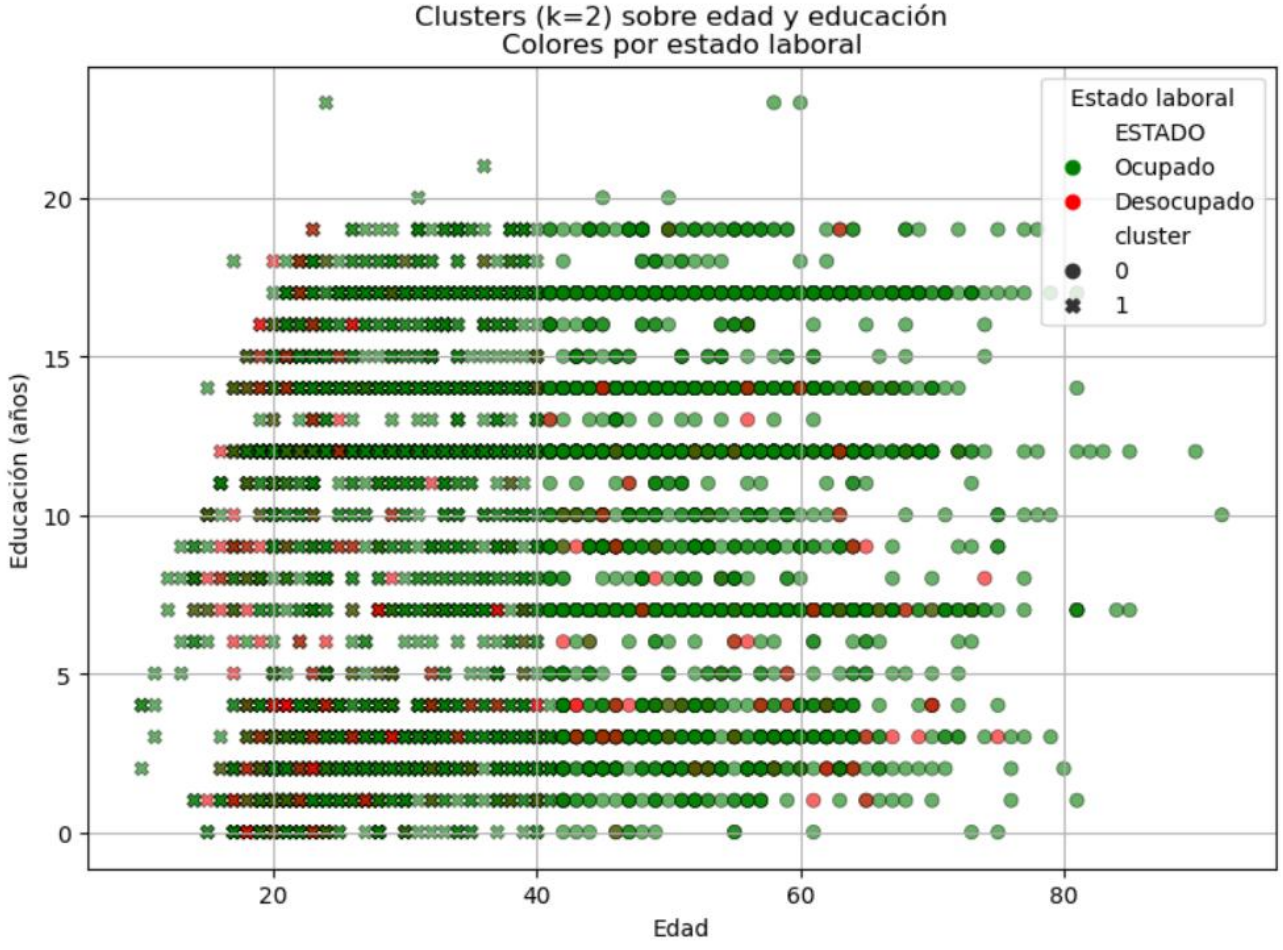
Figura 9 - Región seleccionada: Gran Buenos Aires



En la figura 9 se ve que los resultados muestran cómo el algoritmo K-means divide la población en grupos progresivamente más finos. Con k = 2, se observa una segmentación binaria clara, posiblemente entre niveles socioeconómicos. Con k = 4, aparecen patrones más detallados, como subgrupos por edad o jornada laboral. Con k = 10, se distinguen múltiples nichos que podrían capturar combinaciones específicas de educación, horas trabajadas y salarios. Estos resultados pueden usarse para construir perfiles sociales más precisos en estudios descriptivos o modelos predictivos.

5b. Gráfico edad y educ

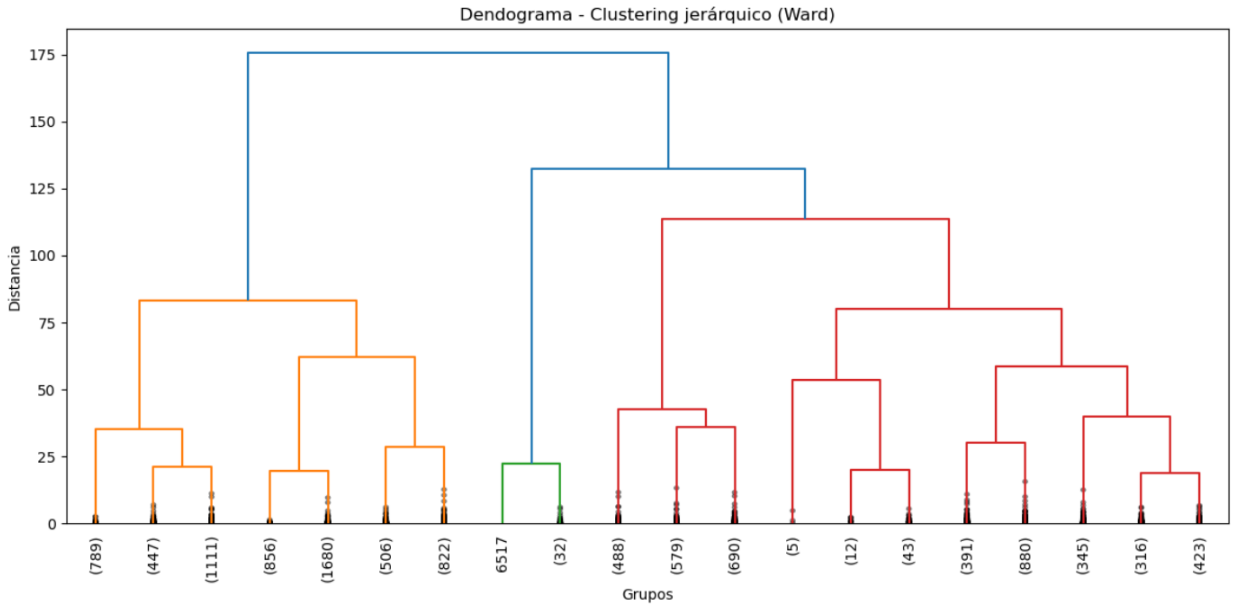
Figura 10 - Región seleccionada: Gran Buenos Aires



En la figura 10, el gráfico evidencia que el clustering no supervisado con K-means (k=2) utilizando solo edad y años de educación no logra distinguir correctamente entre ocupados y desocupados. Aunque hay cierta agrupación estructural, se observa un fuerte solapamiento entre los dos estados laborales dentro de ambos clusters. Esto sugiere que para predecir la situación laboral con mayor precisión se requiere incluir más variables relevantes, como salario, tipo de empleo, y horas trabajadas. En conclusión, el algoritmo descubre dimensiones latentes, pero no coincide perfectamente con la estructura real del mercado laboral.

6. Dendograma

Figura 11 - Región seleccionada: Gran Buenos Aires



En la figura 11, se realizó un análisis de clustering jerárquico aglomerativo utilizando el método de Ward sobre las variables estandarizadas: edad, edad al cuadrado, años de educación, salario semanal y horas trabajadas. El dendograma resultante muestra la secuencia de fusiones entre observaciones similares y permite visualizar la estructura de los datos sin necesidad de fijar a priori el número de grupos. Al cortar el árbol a distintas alturas, pueden identificarse agrupamientos consistentes, lo que sugiere la presencia de subgrupos en la población con patrones comunes en edad, inserción laboral y formación. Este enfoque complementa los resultados del análisis de componentes principales y del clustering K-means, reforzando la segmentación detectada entre individuos con trayectorias socioeconómicas distintas.

Conclusión

El trabajo reveló que la población ocupada se concentra en edades de 20-55 años con mayor desocupación juvenil (18-30 años), mientras que la educación promedio (7.75 años) no alcanza el secundario completo, evidenciando desafíos en capital humano. Los salarios mostraron alta desigualdad (distribución asimétrica) y las horas trabajadas heterogeneidad (50% en jornadas de 40 horas, pero con casos extremos de hasta 100 horas). El PCA destacó que el 68% de la varianza se explica por ingresos/educación (PC1) y edad/horas trabajadas (PC2), mientras que el clustering (k-means y jerárquico) identificó segmentos socioeconómicos pero no diferenció eficientemente ocupados de desocupados, sugiriendo la necesidad de incluir más variables. Las mejoras en ocupación (40.3% a 45.7%) y reducción de desocupación (6.9% a 4.4%) entre 2004-2024 reflejan progreso, pero persisten brechas que requieren políticas focalizadas en educación, formalización laboral y apoyo a jóvenes.

Referencias

- *INDEC (2025). Encuesta Permanente de Hogares: Metodología. Disponible en: https://www.indec.gob.ar/*
- *Schwabish, J. A. (2014). An economist's guide to visualizing data. Journal of Economic Perspectives, 28(1), 209-23*