MACHINE LEARNING

LAURA GARCÍA GONZÁLEZ Y LUCÍA MARTÍNEZ MIRAMONTES

ÍNDICE

1. INTRODUCCIÓN
2. PREPROCESAMIENTO
3. HIERARCHICAL CLUSTERING
4. PARTITIONAL CLUSTERING
5. DBSCAN
6. GAUSSIAN MIXTURE MODELS
7. CONCLUSIÓN

INTRODUCCIÓN

- Aplicación de técnicas de clustering sobre un conjunto de datos socioeconómicos para hallar agrupaciones naturales
- Preprocesamiento exhaustivo para igualar el peso de las variables
- Evaluación del efecto de los distintos parámetros

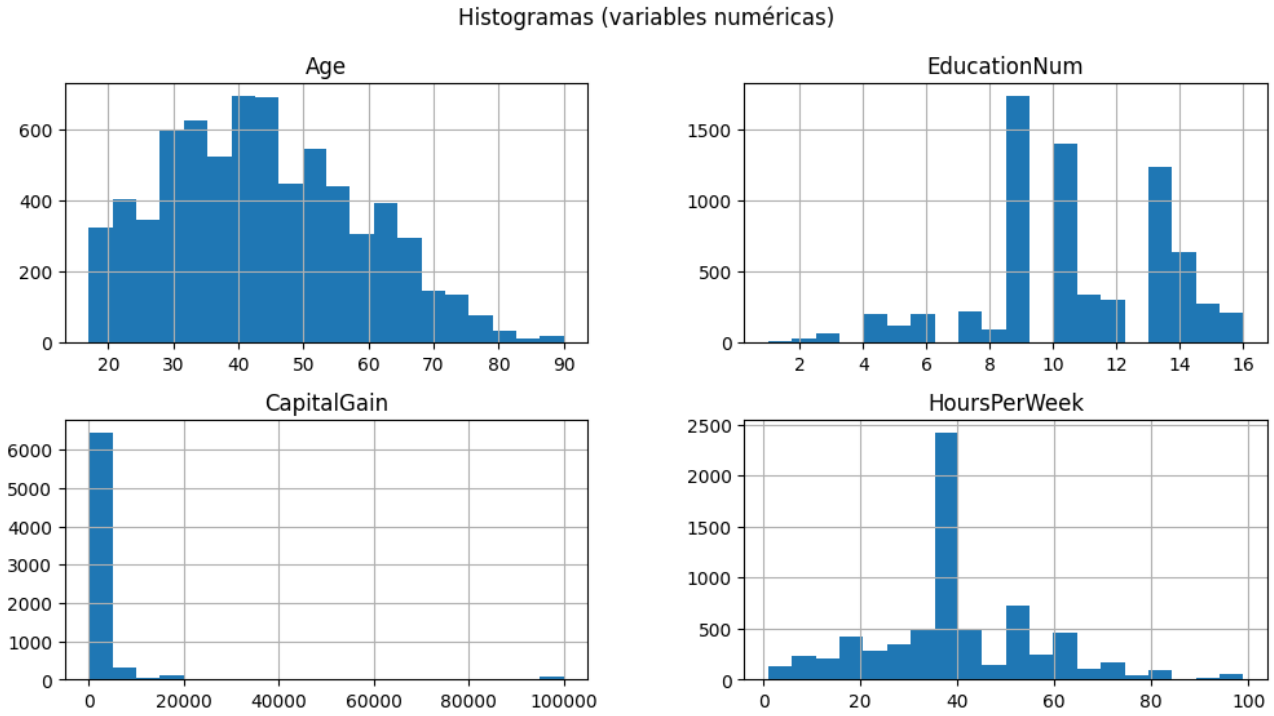
PREPROCESAMIENTO

- **Consistencia del formato:**
  - Datos estandarizados, sin inconsistencias de mayúsculas o espacios.
  - Revisados los valores numéricos, sin valores fuera de rango.

- **Valores faltantes:**
  - Ninguna columna o fila tiene datos nulos → no se requirió imputación.

- **Duplicados:**
  - Filas duplicadas eliminadas para evitar redundancia y optimizar el clustering.

OUTLIERS

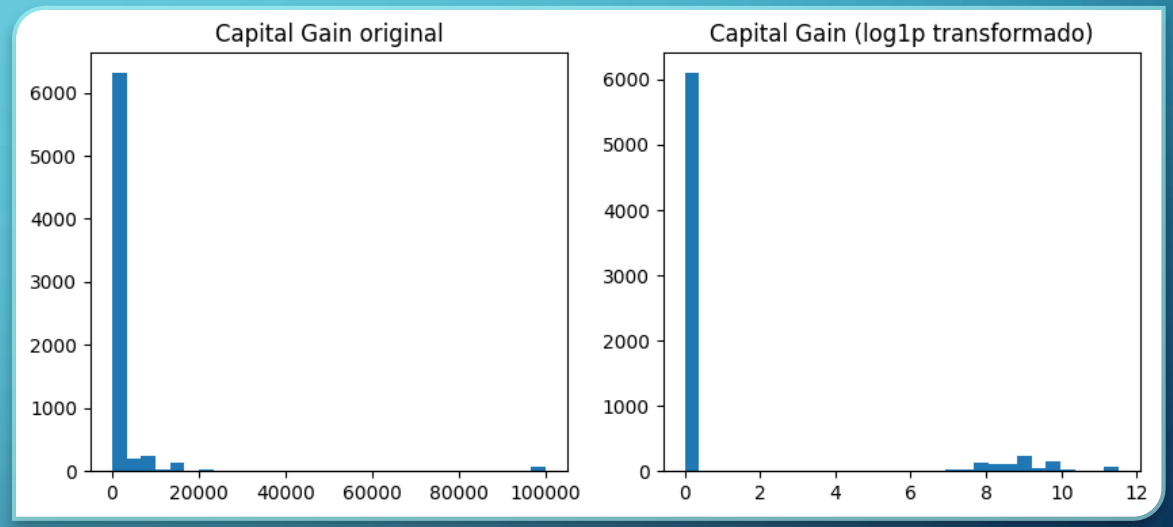

Histogramas (variables numéricas)
Age
EducationNum
CapitalGain
HoursPerWeek

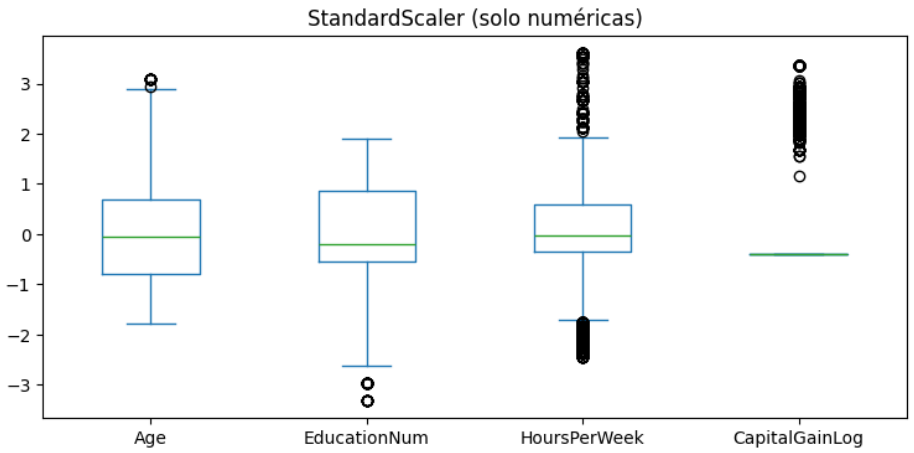
CAPITAL GAIN

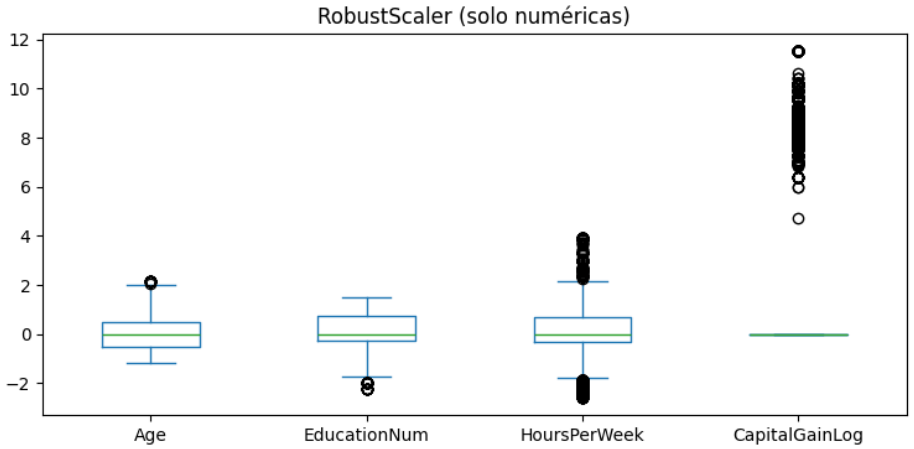
•Distribución muy sesgada: muchos ceros y valores extremos.

•Se aplicó **transformación logarítmica**:
- •Reduce el efecto de los valores extremos.
- •Maneja ceros de forma segura.
- •Hace la variable más homogénea.

•Aunque sigue siendo asimétrica, el rango se reduce, logrando una distribución más adecuada para el clustering.







VARIABLES REDUNDANTES Y ESCALADO

- **Correlación Education vs EducationNum:**
  - Codificación inicial sin orden → correlación débil (0.383).
  - Orden correcto aplicado → correlación muy alta (0.995).
  - Conclusión: **Education es redundante** respecto a EducationNum.
- **Escalado de variables numéricas:**
  - Se evaluó para evitar que variables con rangos mayores dominen el clustering.
  - Métodos probados: **StandardScaler** y **RobustScaler**.

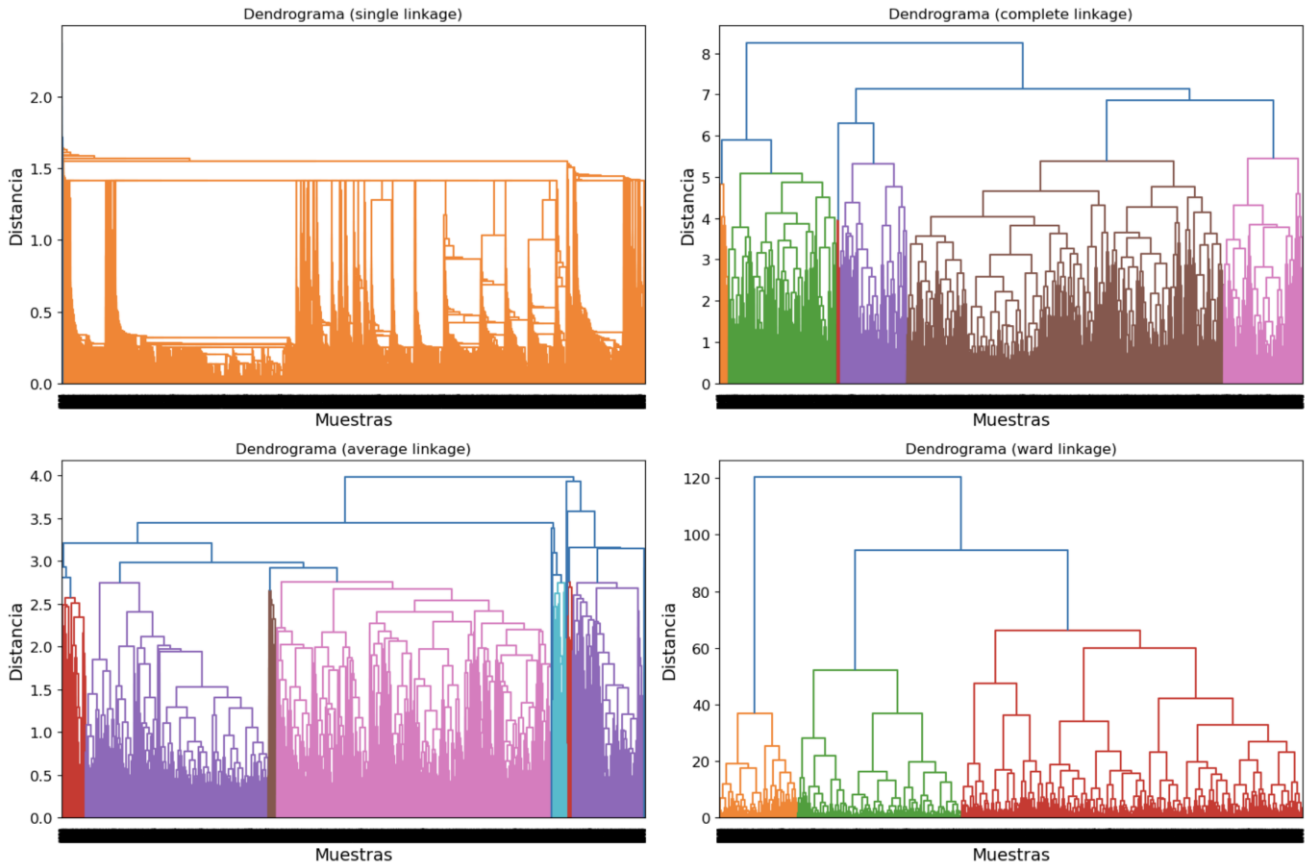
VARIABLES CATEGÓRICAS

- **CapitalGain:** Posible descomposición en dos componentes, variable con muchos ceros (86.5%) y valores extremoss:
  - **Binaria:** indica si hay ganancias o no.
  - **Logarítmica:** valores de ganancias para reducir asimetría
- Idea descartada, correlación del 0.99 (coeficiente de Pearson)

- **Columnas categóricas:** Gender, MaritalStatus, Relationship (Education eliminada por redundancia).

- **Codificación:**
  - Se descartó OrdinalEncoder (variables nominales, sin orden natural).
  - **Gender:** One-Hot Encoding, eliminando una columna por redundancia.
  - **MaritalStatus y Relationship:** One-Hot Encoding, cada categoría como columna binaria.

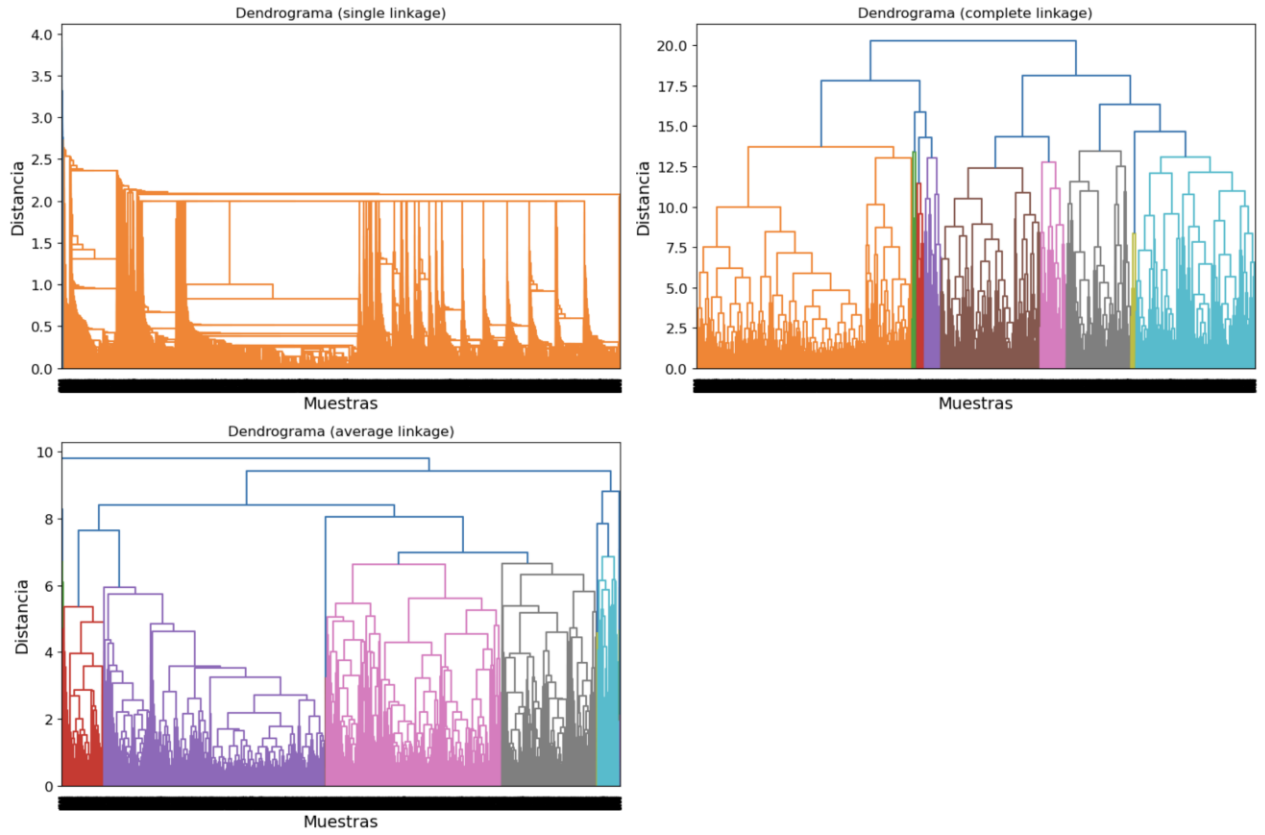
- **Consideraciones:**
  - Aumenta dimensionalidad.
  - Permite aplicar clustering correctamente y obtener resultados comparables.



HIERARCHICAL CLUSTERING - EUCLIDEA

- Se escogió el **dendrograma Ward** como el más adecuado.
- Ward mostro un árbol más estable, saltos más largos, y fusión gradual, indicando agrupamientos naturales.
- Conclusión: número de clusters óptimo entre 2 **y 4**.

HIERARCHICAL CLUSTERING - MANHATTAN



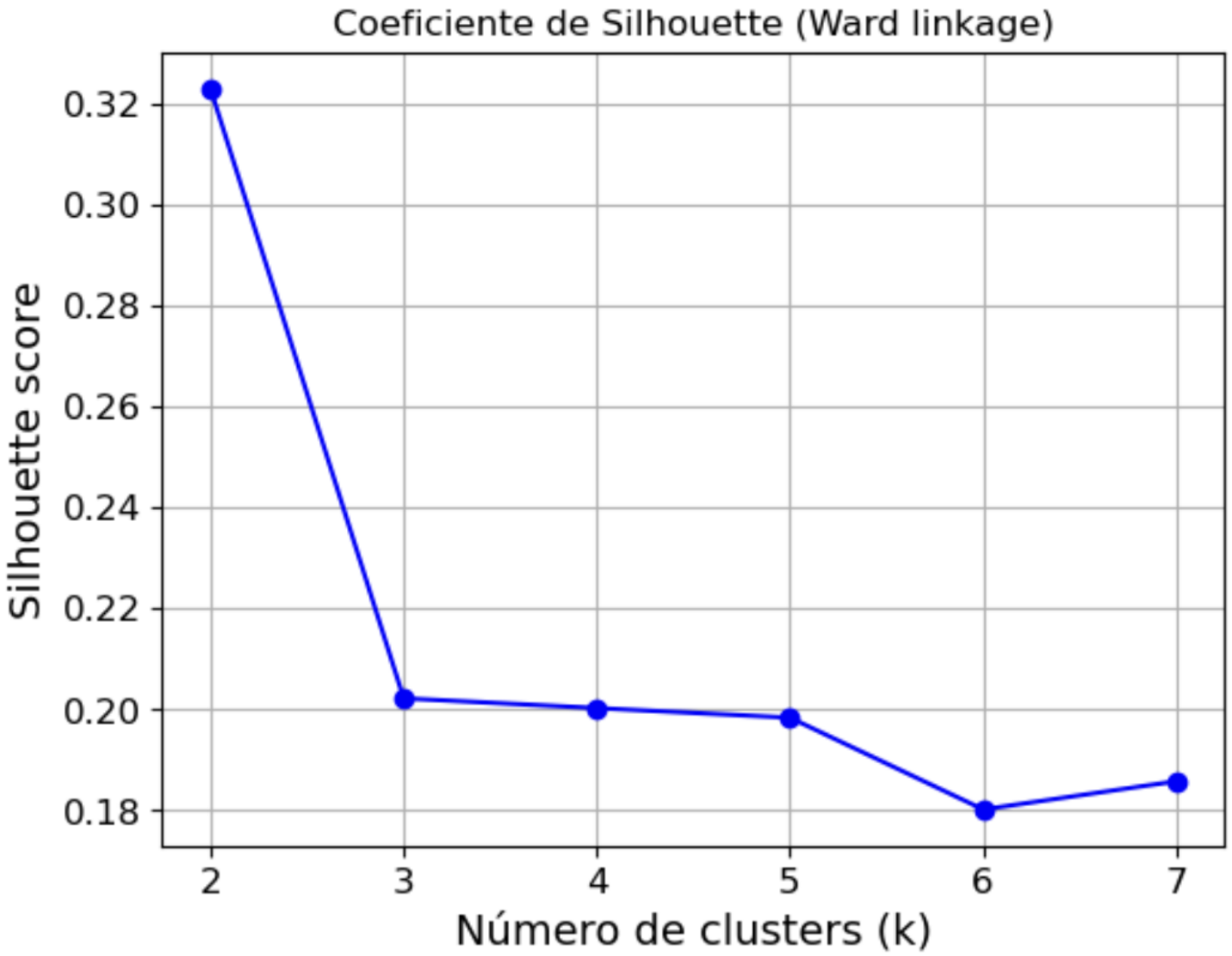
- Euclídea → saltos claros en el dendrograma, permitiendo identificar clusters naturales.

- Manhattan → dendrograma poco informativo, sin agrupaciones intermedias claras.

HIERARCHICAL CLUSTERING - SILHOUTTE SCORE

- Pico claro en K = 2, los datos tienen una separación natural en dos perfiles.

- Estructura jerárquica débil, los coeficientes indican alto solapamiento de los clusters en general



valores numéricos

cluster	Age	EducationNum	HoursPerWeek	CapitalGainLog
0	49.25	9.98	42.59	0.00
1	48.17	11.42	43.34	8.91
2	37.67	10.79	37.94	0.00

valores categóricos

cluster	MaritalStatus	Relationship	Gender
0	Married-civ-spouse	Husband	Male
1	Married-civ-spouse	Husband	Male
2	Never-married	Not-in-family	Female

valores numéricos

cluster	Age	EducationNum	HoursPerWeek	CapitalGainLog
0	43.04	10.41	40.09	0.00
1	48.17	11.42	43.34	8.91

valores categóricos

cluster	MaritalStatus	Relationship	Gender
0	Married-civ-spouse	Husband	Male
1	Married-civ-spouse	Husband	Male

HIERARCHICAL CLUSTERING – CLUSTERINGS FINALES

- Para K = 2 y K = 3: variable determinante es CapitalGainLog -> explica que haya una division en dos de los datos
- Variables cuantitativas a penas tienen peso (se mantienen constantes a lo largo de los distintos clusterings)

HIERARCHICAL CLUSTERING-CONCLUSIÓN

- **CapitalGain** domina la formación de clusters, separando claramente a quienes tienen ganancias de los que no.

- Edad, educación, horas trabajadas y género generan subgrupos que se fusionan al reducir k.

- El clustering jerárquico no es arbitrario(se observaron más valores de k) pero las agrupaciones son mixtas y solapadas.

- **k = 4** podría capturar mejor subgrupos importantes, como mujeres jóvenes, no casadas y con alta educación.

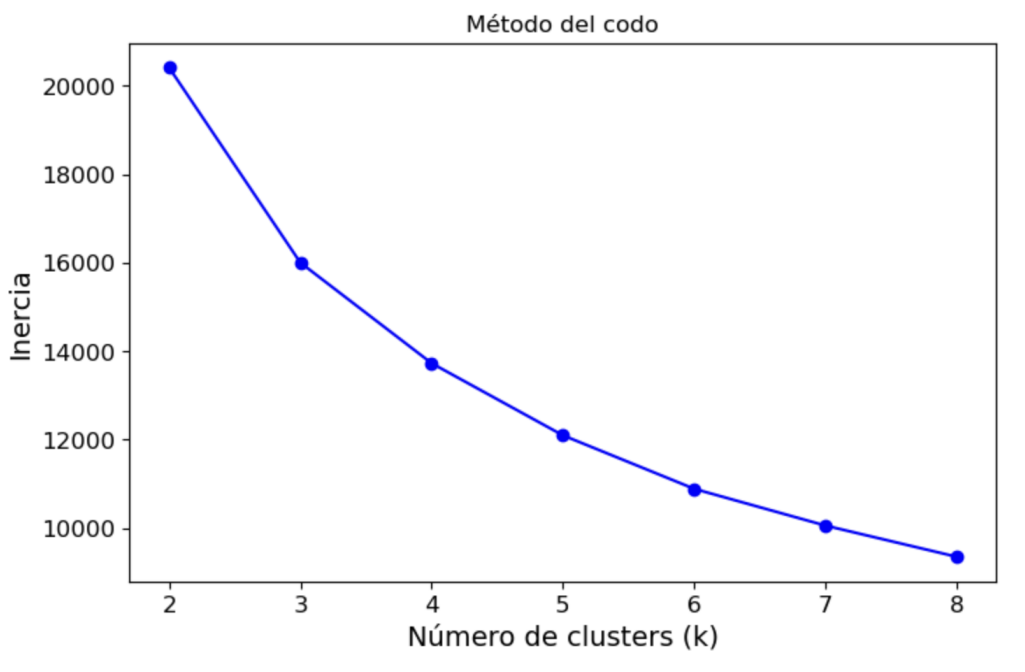
PARTITIONAL CLUSTERING – REDUCCIÓN DE DIMENSIONALIDAD

- Reducción de dimensionalidad, reducir dominancia de las variables categóricas por cantidad excesiva de dummies y CapitalGain en el estudio de distancias
  - PCA - valores numéricos escalados (n_components = 3)
  - MCA – variables categóricas (n_components = 3)

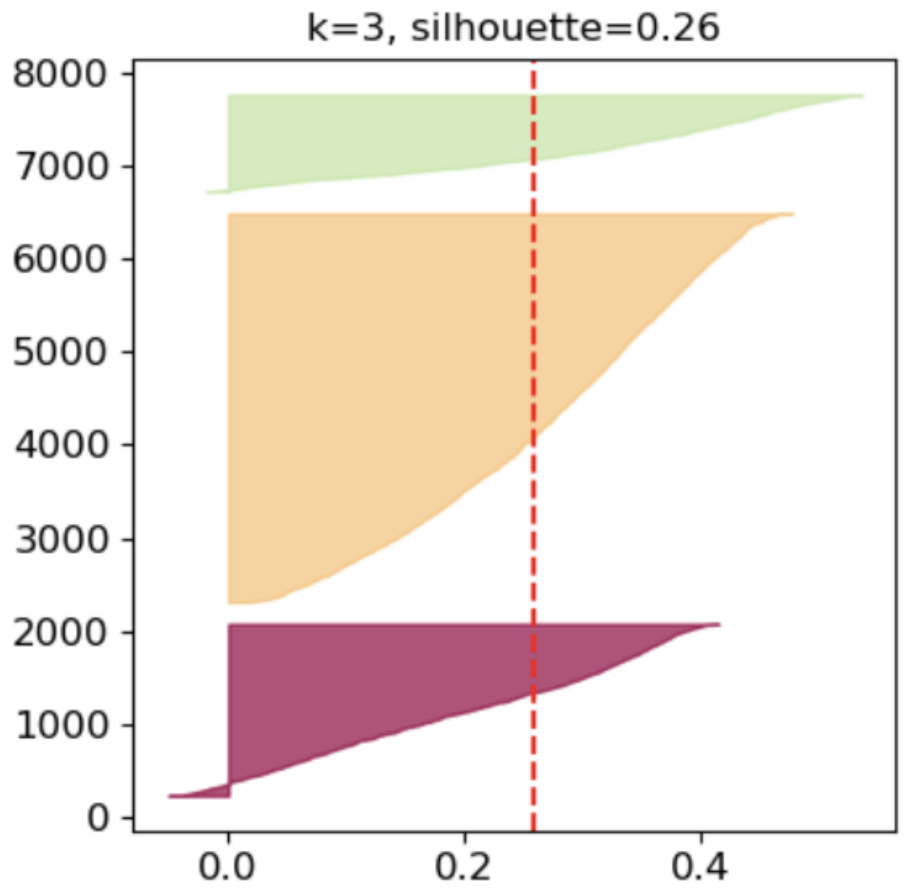
  Posibles correspondencias entre categóricas (Male-Husband, Female-Wife, "Husband-Married", etc.)

PARTITIONAL CLUSTERING - KMEANS

INERCIA (MÉTODO DE CODO)



COEFICIENTES SILHOUTTE



PARTITIONAL CLUSTERING – SILHOUTTE SCORES

- Valores positivos en casi su totalidad
- Bandas relativamente anchas y compactas
- El Segundo cluster recoge la mayoría de los datos
- Coeficientes bajos pero positivos, propios de datos Socioeconómicos
- Refuerzan la elección de k = 3

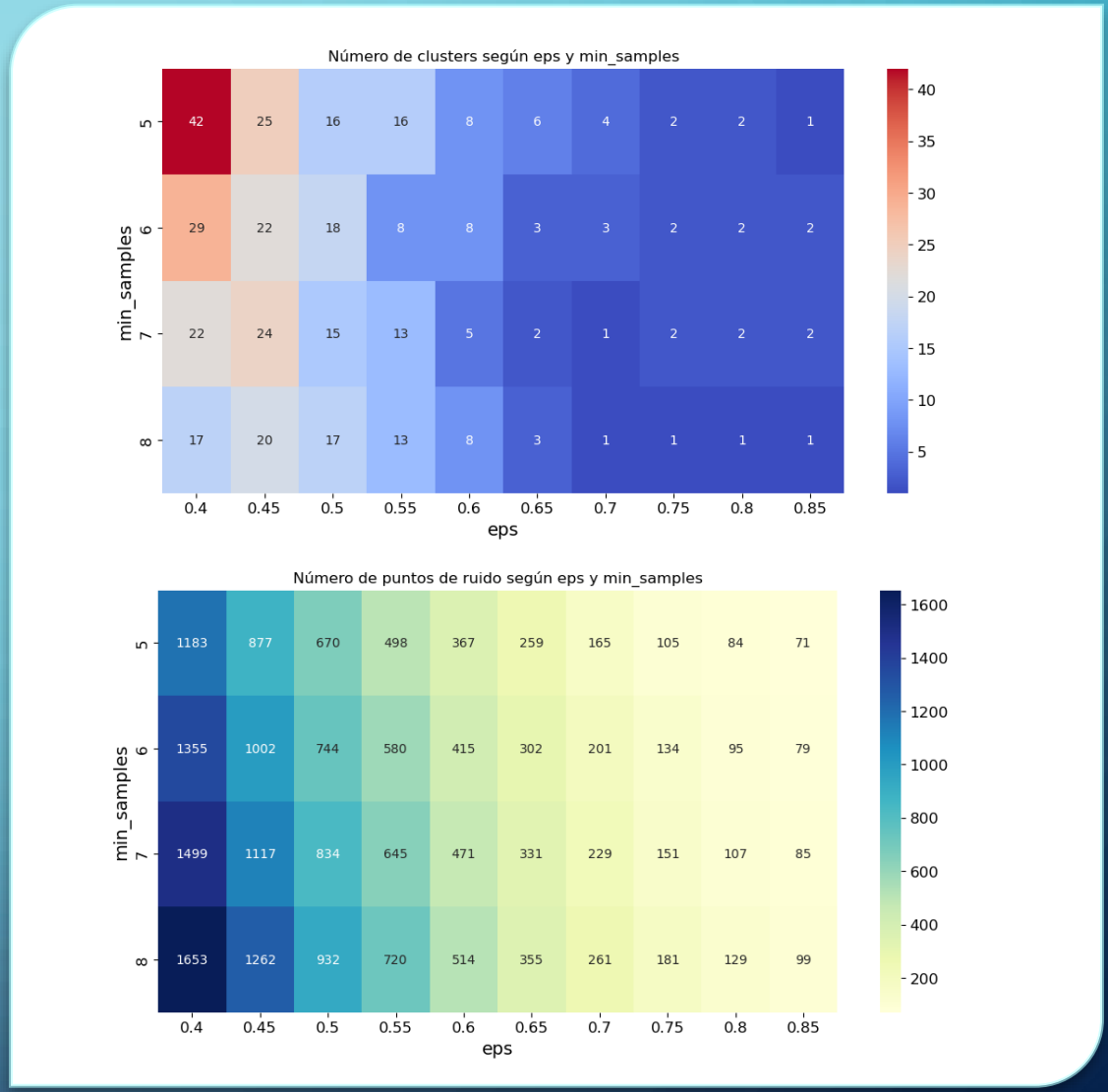
PARTITIONAL CLUSTERING – RESULTS

	MaritalStatus_top	MaritalStatus_pct	Relationship_top	Relationship_pct	Gender_top	Gender_pct
0	Married-civ-spouse	0.72	Husband	0.66	Male	0.79
1	Married-civ-spouse	0.38	Husband	0.3	Male	0.62
2	Married-civ-spouse	0.53	Husband	0.45	Male	0.62

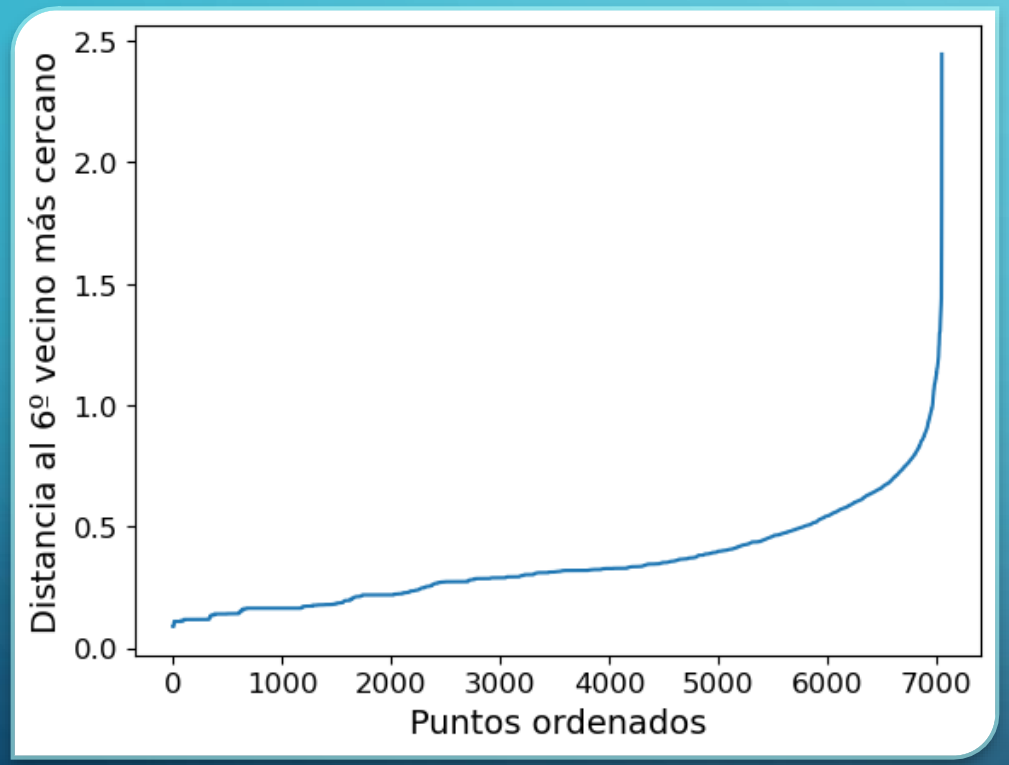
	n_points	mean_age	mean_hours	mean_education	mean_gain
0	1051.0	50.46	46.86	12.27	7.41
1	4190.0	35.79	43.42	11.08	0.04
2	1814.0	58.17	30.20	8.33	0.29

DBSCAN

- Parámetros eps y min_samples
- Evaluación preliminar sobre un rango de valores para determinar min_samples fijo
- Entre 5 y 2*d (d = dimensionalidad de los datos)
- Se eligió min_samples = 7
  - Número de clusters razonable y estable



DBSCAN – REGLA DEL "CODO"

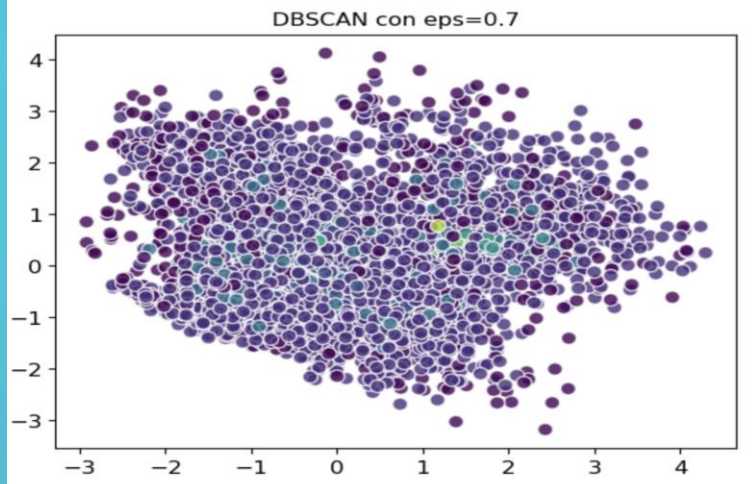


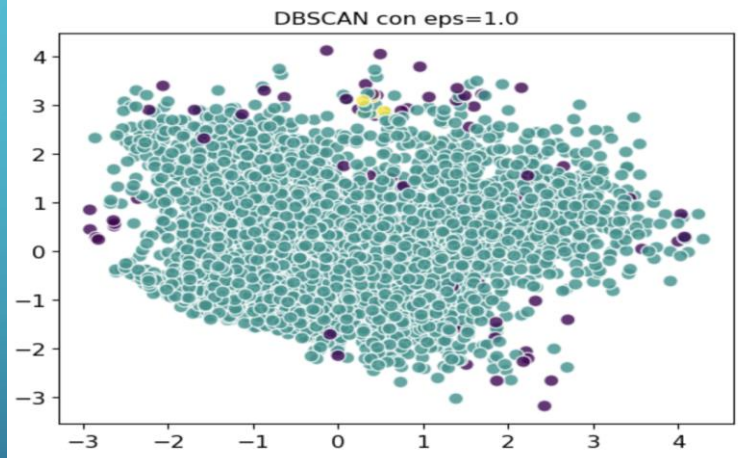
- Buscamos los valores en los que cambia el crecimiento lento a drástico
- Rango entre 0.6-1.2

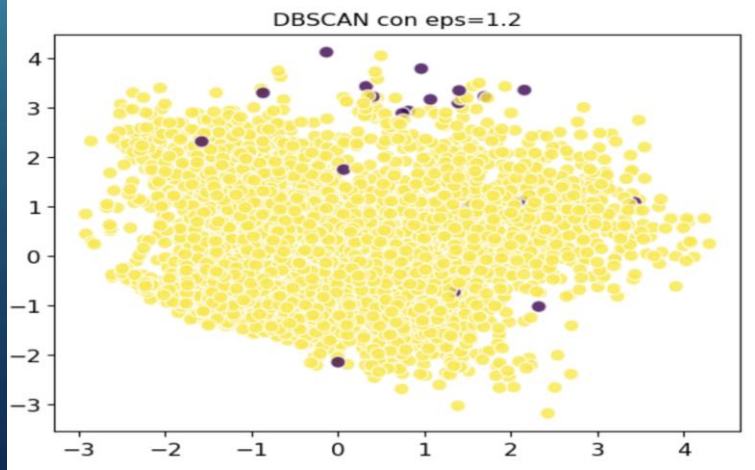
DBSCAN – DISTINTOS EPS

- Se probaron distintos eps dentro del rango
- Visualización en 2D (se marcan los outliers detectados)
- Evaluación cuantitativa con coeficientes silhoutte
- Se escogió eps = 1.0 (mejor score y cantidad de clusters adecuada)

	eps	clusters	noise_ratio	silhouette
0	0.6	25	14.755493	-0.216297
1	0.7	8	9.142452	-0.187598
2	0.8	6	4.833451	-0.039418
3	0.9	2	2.381290	0.192445
4	1.0	2	1.133948	0.245393
5	1.1	1	0.680369	NaN
6	1.2	1	0.326010	NaN







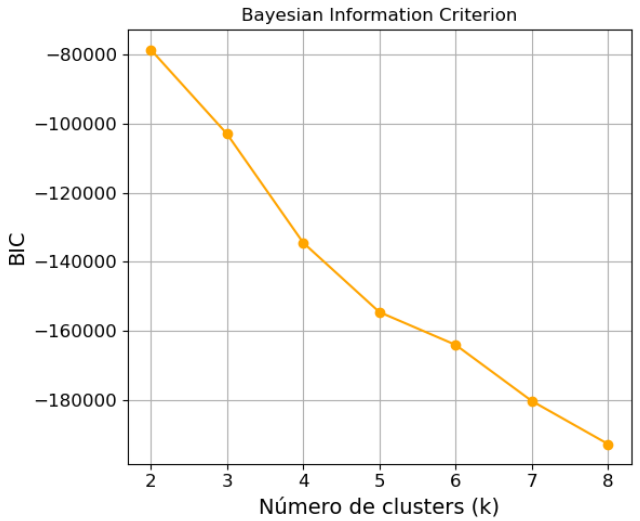
DBSCAN – CLUSTERING FINAL

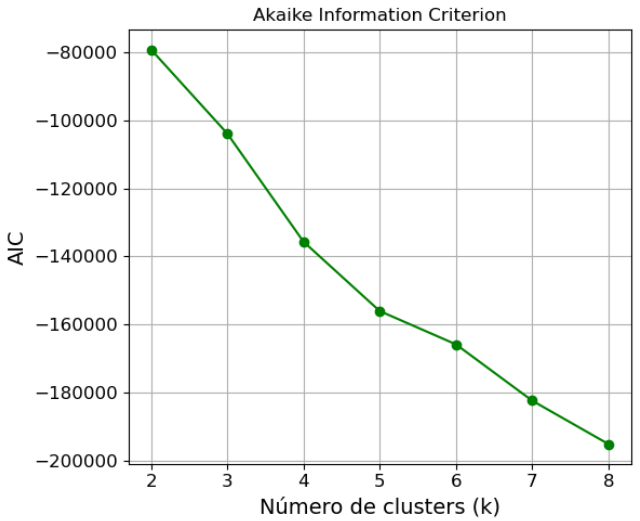
- DBSCAN no consiguió identificar agrupamientos significativos e interpretables
- Un cluster de solo 6 individuos caracterizado por mujeres de mayor edad viudas
- No representa grupos naturales
- DBSCAN busca zonas de alta densidad claramente separadas, nuestros datos contienen mucho solapamiento y forman una gran nube

	MaritalStatus_top	MaritalStatus_pct	Relationship_top	Relationship_pct	Gender_top	Gender_pct
-1	Never-married	0.32	Not-in-family	0.44	Female	0.55
0	Married-civ-spouse	0.48	Husband	0.4	Male	0.65
1	Widowed	0.83	Not-in-family	0.83	Female	1.0

	n_points	mean_age	mean_hours	mean_education	mean_gain
-1	80.0	57.71	52.05	8.62	3.92
0	6969.0	43.55	40.42	10.57	1.17
1	6.0	68.50	17.83	9.17	8.12

GAUSSIAN MIXTURE MODELS





BIC y AIC disminuyen con k, lo que muestra que modelos con más componentes ajustan mejor, pero añaden complejidad.

	k	Silhouette	BIC	AIC
0	2	0.105205	-78508.459686	-79119.132462
1	3	0.083920	-102914.143979	-103833.583889
2	4	0.023485	-134431.041758	-135659.248802
3	5	0.040601	-154533.203453	-156070.177631
4	6	0.018695	-164022.203756	-165867.945068
5	7	0.016301	-180284.005757	-182438.514203
6	8	-0.004548	-192741.145031	-195204.420610

GMM – CLUSTERING FINAL

	MaritalStatus_top	MaritalStatus_pct	Relationship_top	Relationship_pct	Gender_top	Gender_pct
0	Never-married	0.61	Not-in-family	0.51	Male	0.5
1	Married-civ-spouse	0.43	Wife	0.39	Female	0.79
2	Married-civ-spouse	1.0	Husband	1.0	Male	1.0

	n_points	mean_age	mean_hours	mean_education	mean_gain
0	2979.0	36.82	39.31	10.56	0.68
1	1310.0	48.27	35.36	10.14	0.93
2	2766.0	49.02	44.30	10.73	1.90

- **Cluster 0:** personas jóvenes, jornada moderada, menor ganancia de capital, solteros.
- **Cluster 1:** personas mayores, casadas, mujeres con rol de esposa.
- **Cluster 2:** personas mayores, casadas, hombres con rol de esposo

CONCLUSIONES

•Escalado de variables esencial para evitar que rangos grandes (p.ej., *CapitalGain*) dominen la distancia.

•PCA y MCA equilibran variables numéricas y categóricas, aunque pueden perder diferencias finas; posible mejora: discretizar variables continuas en rangos comparables.

•Explorar distancias o métricas alternativas (Mahalanobis o métricas mixtas) podría mejorar la identificación de clusters y reducir outliers.

•DBSCAN: tamaños limitados.

•GMM y KMeans: divisiones consistentes según edad, género, estado civil y ganancias.

•En conjunto, el dataset presenta **3 clusters naturales**, y combinar métodos facilita detectar outliers y entender la estructura poblacional.

THANK YOU FOR YOUR ATTENTION