APPLIED PROJECT: TOPIC 3

MACHINE LEARNING

DATE: 06/10/2025

Authors:
Laura García González,
laura.gagonzalez@alumnos.upm.es, 3° GCDIA

Lucia Martínez miramontes,
luciammiramentes@alumnos.upm.es, 3° GCDIA

1. Introducción

El objetivo de este proyecto es aplicar técnicas de clustering sobre un conjunto de datos socioeconómicos para descubrir posibles grupos naturales de estos individuos, evaluando y comparando cómo afectan los distintos algoritmos y parámetros a los resultados. Las características recogidas en dichos datos consisten en la edad, nivel educativo, estado civil, relación familiar, género, ingresos de capital y horas trabajadas por semana. Por tanto, se trabaja tanto con datos numéricos (algunos ordinales), como categóricos.

El proceso comenzó con un análisis exploratorio para hallar posibles inconsistencias y un preprocesamiento de los datos, que incluyó la detección de valores faltantes, la codificación de variables categóricas y el escalado de las numéricas. A continuación, se aplicaron distintos algoritmos de clustering (clustering jerárquico, K-means, DBSCAN y modelos de mezcla gaussiana (GMM)) para comparar sus resultados y comprender cómo cada método define los límites entre grupos. El fin fue encontrar el clustering natural, es decir, el que consideramos que fue el más adecuado para nuestros datos.

Los resultados mostraron que los datos tienden a agruparse principalmente en dos o tres perfiles bien diferenciados, en los que se diferencian primordialmente por ganancias de capital, género y estado civil, y edad. Vimos como ciertos métodos lograban hallar estructuras con sentido semántico mientras que DBSCAN identificaba demasiados datos reales como outliers y no era capaz de hallar agrupaciones cohesivas.

2. Método

El proyecto siguió un flujo de trabajo estructurado para analizar los distintos agrupamientos presentes en el conjunto de datos. En primer lugar, se realizó un preprocesamiento exhaustivo, evaluando distintas estrategias de escalado y su impacto en los resultados. Los datos atípicos hallados fueron mantenidos pues consideramos que estos eran datos que reflejaban la realidad y no errores. Se realizó la estandarización de las variables numéricas, usando StandardScaler() para mantener un rango equitativo. A mayores se evaluó la codificación de las variables categóricas mediante diferentes técnicas.

Posteriormente, se aplicaron diversos métodos de clustering. Para cada uno se fueron probando distintos valores de sus parámetros y se concluyó que número de clusters era el indicado a partir de los resultados. Comparamos la calidad de los grupos obtenidos y su interpretabilidad.

Mantuvimos un método experimental, basándonos a menudo en prueba y error, pero no de manera arbitraria. De nuestros resultados pudimos aprender más información acerca de nuestros datos y su distribución, y sobre todo la aparente dominancia de las distintas variables sobre el

funcionamiento de los algoritmos. Con cada prueba podemos modificar nuestro procedimiento, ya sea el preprocesamiento o los inputs, para obtener mejores resultados.

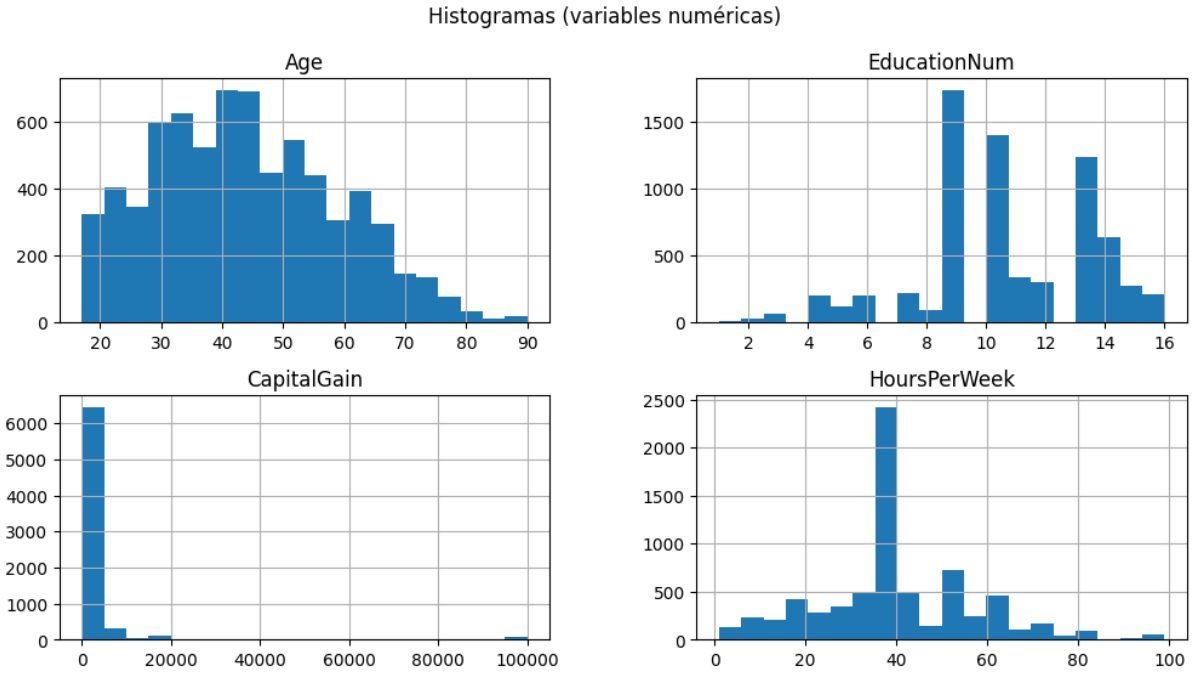
2.1 Preprocesamiento

El preprocesamiento fue un paso crucial en esta práctica, pues los datos obtenidos contenían características complejas y poco manejables que nos trajeron varias dificultades y en sí afectaron al clustering.

Primeramente se realizó una limpieza básica del archivo CSV, como fue la eliminación de espacios y estandarización del formato, evitando errores en herramientas como OneHotEncoder o gráficos basados en conteos. Por último, se revisaron los datos numéricos para detectar valores fuera de rango, como negativos o excesivamente grandes. No se encontraron inconsistencias, por lo que no fue necesaria ninguna corrección y se da por finalizada la estandarización del formato.

Se procedió a analizar la presencia de filas duplicadas en el conjunto de datos. Este paso es relevante, ya que los registros idénticos no aportan información adicional y podrían generar redundancia en el análisis. Aunque algunas filas podrían corresponder a personas diferentes que comparten exactamente las mismas características en las variables, se decidió eliminar los duplicados, ya que afectan mínimamente al cálculo de disimilitud en el clustering y por tanto permiten optimizar el cómputo sin alterar la estructura de los grupos obtenidos.

Un paso significativo fue el análisis preliminar de valores atípicos (outliers) con el objetivo de identificar posibles observaciones extremas que pudieran distorsionar los resultados del clustering. Los valores atípicos se estudiaron principalmente en las variables numéricas (Age, CapitalGain y HoursPerWeek), ya que estas pueden presentar distribuciones con colas largas o valores extremos.

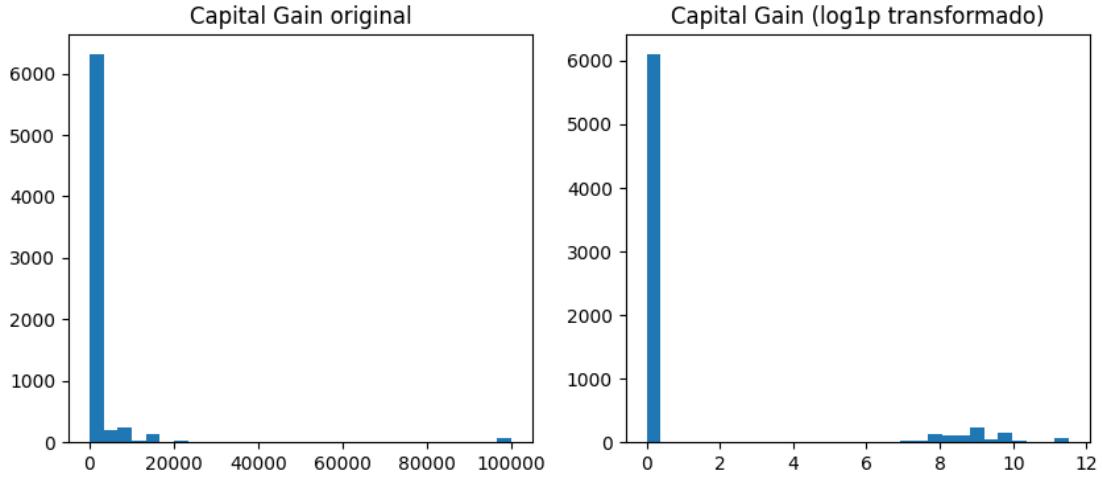


Se observa lo siguiente:

- Age: Distribución muy simétrica con pocos valores atípicos según el IQR.
- EducationNum: Rango estrecho (aprox. 9–14), con algunos valores aislados por debajo de 3 considerados outliers.
- CapitalGain: Distribución muy asimétrica y dispersa, con muchos ceros; los pocos valores distintos se consideran outliers, llegando hasta 99.999.
- HoursPerWeek: Distribución cercana a la normal, con varios outliers en ambos extremos; valores extremos altos (como 99 horas/semana) parecen poco realistas.

Muchos de estos valores representan datos reales que reflejan grupos minoritarios con valor interpretativo y contribuyen a la representación de la realidad. Eliminarlos modificaría la variabilidad y asimetría del dataset. Además, algunos algoritmos de clustering, como DBSCAN, ya están diseñados para manejar este tipo de valores como ruido.

En este estudio, CapitalGain presenta una distribución muy sesgada, con muchos ceros y algunos valores extremadamente altos. Para reducir la asimetría y evitar que los valores extremos dominen el clustering, se aplicó una transformación logarítmica, que además permite manejar los ceros de forma segura y hace que la variable sea más homogénea.

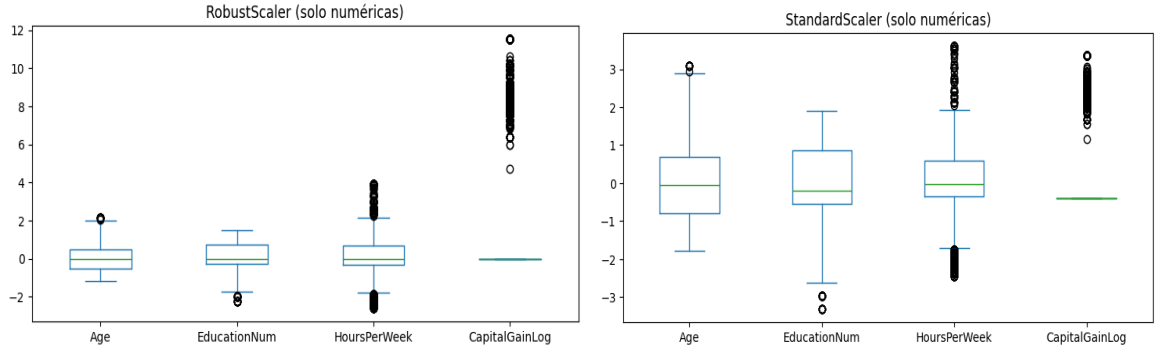


Vemos que, aún manteniéndose la distribución asimétrica, debido a que la absoluta mayoría de los datos son 0, el rango es mucho menor y por tanto se ha logrado adquirir una distribución más adecuada para el alcance de dicha característica.

A mayores, se estudió la correlación entre variables para poder encontrar posibles características redundantes, confirmamos que, efectivamente, la columna EducationNum era una representación numérica de Education, pasando esta última a numérica ordinal con OrdinalEncoder fuimos capaces de calcular el coeficiente de Pearson de este par, obteniendo como resultado 0.995, por lo que decidimos eliminar Education, no se encontraron más correlaciones en las variables numéricas.

Antes de aplicar los algoritmos de clustering, se evaluó la necesidad de escalar las variables numéricas para asegurar que ninguna de ellas dominará el análisis debido a sus rangos distintos, se tuvo especial consideración con CapitalGain por estar “inundada” de ceros (hasta un 86.49%).

Dado que no se eliminaron los datos atípicos, se consideró que RobustScaler sería un método de escalado más adecuado al StandardScaler, se muestra la comparación en la siguiente figura.



Nos fijamos en CapitalGain, vemos en estos resultados que RobustScaler parece ser demasiado agresivo para nuestros datos ya transformados por aplicación de logaritmos, esto se debe de nuevo a la cantidad de ceros, pues, realizando un resumen estadístico tanto la media, la mediana, los cuartiles, etc. se encuentran en 0. Al usar este escalado los datos extremos de esta variable dominarán nuestros clusterings. Por tanto, a pesar de que sí hemos mantenido los valores atípicos, el escalado más adecuado para esta variable resulta ser StandarScaler.

Decidimos además, que dado que la escala resultante para el resto de variables permanecía prácticamente constante con ambos escaladores, usaríamos StandardScaler para todos, con ánimo de mantener unicidad.

Este exceso de ceros tiene implicaciones importantes: al tratar la variable como continua, se está considerando a personas sin ganancias junto con aquellas que sí las tienen, lo que genera un grupo extremadamente compacto de ceros y valores positivos muy dispersos. Esto puede causar que muchos algoritmos de clustering consideren los valores positivos como outliers o generen agrupaciones poco realistas.

De manera adicional a la aplicación de logaritmos, consideras descomponer CapitalGain en dos componentes, esta misma y una nueva característica que llamaremos 'HasCapitalGain', binaria, y que indica si una persona tiene o no ganancias.

El objetivo era mejorar la estructura del espacio de distancias que mejorará el clustering. Al separar la existencia de ganancias (información categórica) de su magnitud (información continua suavizada), el modelo puede diferenciar entre individuos con y sin ingresos de capital sin que los valores extremos distorsionen las distancias. Esto genera grupos más equilibrados, donde CapitalGain aporta información relevante pero no desproporcionada.

A pesar de esto, descubrimos que estas dos variables resultaban tener una correlació por coeficiente de Pearson del 0.99, dando a ha entender, que la inundación de 0s en CapitalGainLog

es tán alta que al final, la información que aporta es, en esencia, si la persona tiene o no ganancia (comportamiento binario).

El último paso del preprocesamiento fue la conversión de variables categóricas a numéricas, necesaria para los modelos de clustering. Dado que Education ya fue eliminada, el resto de variables no eran ordinales, por ello concluimos que OneHotEnocder sería la opción más adecuada y fue aplicada sobre las columnas categóricas: Gender, MaritalStatus y Relationship.

Para la realización del preprocesamiento limpio y reproducible se usó ColumnTransformer indicando que transformaciones se realizan sobre qué variables del dataframe.

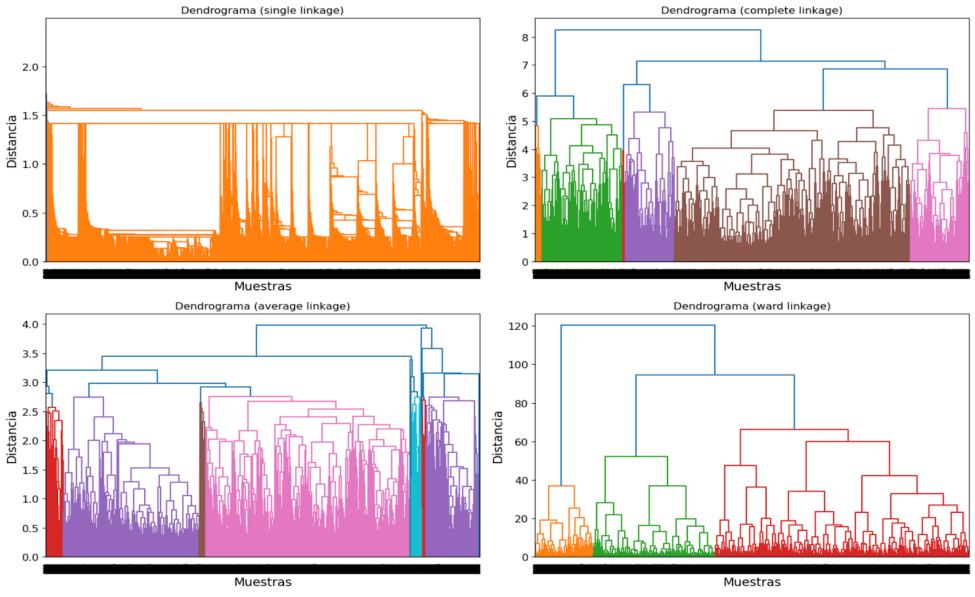
3. Resultados

A continuación se expondrá en detalle los resultados obtenidos, evaluando las diferentes técnicas de clustering y explicando los resultados obtenidos de estas. Finalmente se analizará el clustering elegido como el más óptimo para nuestros datos.

3.1 Clustering Jerárquico

Para analizar la estructura jerárquica de los datos se aplicó el clustering jerárquico aglomerativo, evaluando diferentes métodos de enlace (linkage): single, complete, average y Ward. Cada método determina cómo se calculan las distancias entre clusters al momento de fusionarlos, lo que influye directamente en la forma del dendrograma y en la identificación de grupos naturales.

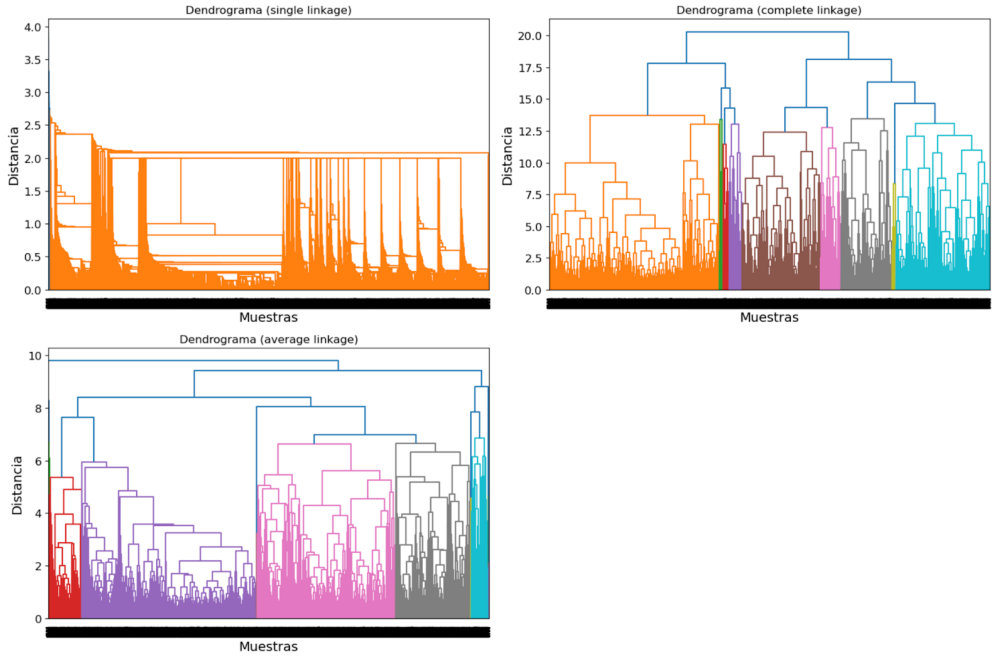
En primer lugar, se construyó un dendrograma para cada tipo de linkage utilizando la distancia euclídea, dado que esta medida es la más adecuada para los métodos jerárquicos basados en varianza, como el linkage Ward, y permite evaluar de forma consistente la similitud entre las observaciones numéricas estandarizadas.



Podemos ver los siguientes resultados:

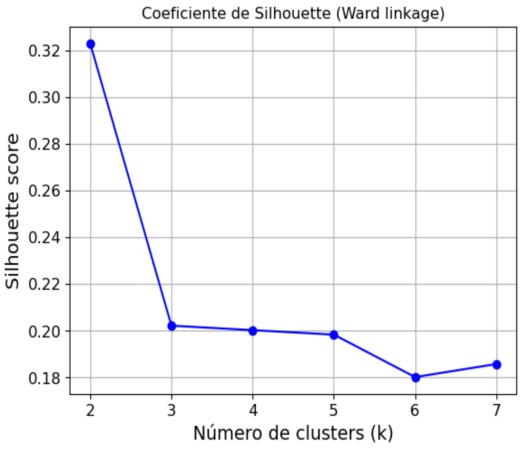
- **Single Linkage:** las instancias han resultado en grupos que se forman sin saltos grandes ya que el rango de distancia llega solo hasta 2. Esto se debe a que con single linkage las fusiones ocurren muy pronto, pues se están teniendo en cuenta las distancias mínimas, por ello vemos largas cadenas en el dendrograma. Nos está mostrando un árbol poco estable. Definitivamente, single linkage no resulta útil para hallar clusters reales.
- **Complete Linkage:** en este vemos una gran mejora respecto al anterior, se ve un árbol mucho más estable. No obstante los saltos son significativamente cortos dando a entender que las agrupaciones no muestran diferencias potentes.
- **Average Linkage:** es un intermedio entre single y complete ya que se basa en medias, por tanto si se ve una jerarquía más clara que para single, pero más alargada y con saltos menos distantes que con complete linkage. En este caso el average linkage no parece ser adecuado.
- **Ward Linkage:** Finalmente, Ward muestra un dendrograma que a nivel visual se ve muy estructurado y equilibrado visualmente, los saltos son más grandes que para los demás tipos pero son pocos, realizando agrupaciones de manera proporcional, las alturas de fusión crecen muy suavemente. Para nuestros datos escalados (estandarizados) es el más coherente.

A mayores se construyeron los dendrogramas utilizando la distancia Manhattan, con el fin de comparar cómo varía la estructura de los clusters al cambiar la métrica de distancia. La distancia Manhattan mide la suma de las diferencias absolutas entre las coordenadas de las observaciones, lo que puede resaltar relaciones diferentes entre los puntos, especialmente en datasets con variables escaladas y dummies. En este caso, no se puede aplicar Ward linkage con la distancia Manhattan porque Ward requiere necesariamente la distancia euclídea.



Decidimos que el mejor método de linkage es ward, por supuesto entonces con la distancia euclídea, pues vemos que con la distancia Manhattan no se ha podido superar la estabilidad y separación lograda en este otro.

Fijándonos en el dendrograma de ward, podemos interpretar que los clusterings más razonables se dan entre 2 y 4 clusterings pues se muestran los más separados (por longitud de sus saltos) y claros. K = 3 parece la opción más natural. Para apoyar nuestras interpretaciones realizamos un estudio de los coeficientes silhoutte para un amplio rango de valores de k.



Se muestra que el verdadero pico esta en 2 clusters, aunque por lo general los resultados de este coeficiente son muy bajos, lo que nos lleva a que hay una estructura débil en los datos, es decir, son mixtos y por tanto hay un alto solapamiento de clusters aún incluso cuando sólo hay 2.

El hecho de que el coeficiente de Silhouette muestre un máximo claro para k=2 y que el dendrograma evidencie un gran salto en la distancia de fusión al pasar de dos a un solo grupo sugiere que los datos presentan una separación natural en dos perfiles principales.

Este comportamiento es coherente con la estructura de las variables empleadas en el clustering, CapitalGainLog, aunque transformada, continúa siendo una variable muy desequilibrada y con alta concentración de valores cero. Esto provoca que una parte sustancial de las observaciones tenga ganancias de capital nulas, mientras que un grupo reducido presenta valores significativamente más altos.

Esta dicotomía genera una separación fuerte entre individuos con y sin ganancias de capital, que dominan la estructura de los clusters. Indagamos en cómo era este clustering óptimo y obtuvimos datos que respaldan justamente esto:

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

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

Es decir, vemos que la agrupación se ha realizado basándose al completo en la obtención o no de ganancias de capital, mientras que las demás variables se mantienen constantes.

Variables como HoursPerWeek, Age y EducationNum aportan variabilidad adicional, pero no suficiente para definir subgrupos claramente separados dentro de los dos principales.

Observamos los datos para k = 3, nuestra otra sospecha:

cluster	Age	EducationNum	HoursPerWeek	CapitalGainLog
0	41.53	10.29	34.92	0.00
1	48.17	11.42	43.34	8.91
2	46.19	10.66	50.89	0.00

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

De estas tablas obtenemos conclusiones valiosas, es evidente que las variables categóricas, en las que podríamos incluir CapitalGain por su comportamiento binario, dominan el clustering.

Estos saltos abruptos (0-1) se traducen a mayores distancias que los rangos continuos que tienen las variables numéricas.

Como conclusión, tanto de manera interpretativa como cuantificada, se entiende que hay una jerarquía de clustering significativa en el contexto de nuestros datos, pues las fusiones no son arbitrarias, sin embargo, estos agrupamientos resultan ser débiles y basarse primordialmente en una única característica. Vimos que incluso al tener k = 5 los valores de las variables continuas difirieron pero no de manera determinante, simplemente como resultado de las distribuciones.

3.2 Clustering Particional

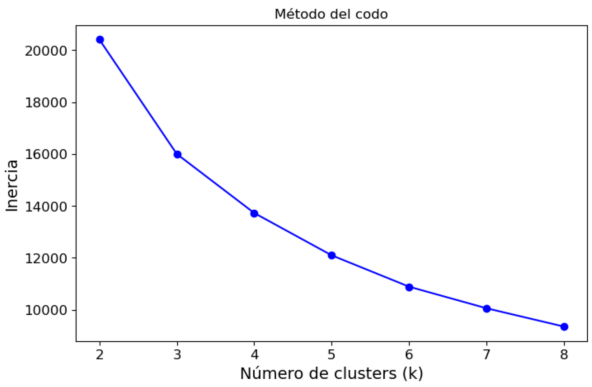
Se realizó el clustering mediante la técnica de K-Means, para saber que valor del parámetro k escoger, se realizaron pruebas con un rango de valores y se evaluaron los resultados de dos maneras, con el estudio de la inercia (método del codo) y los coeficientes de silhoutte (búsqueda de picos.

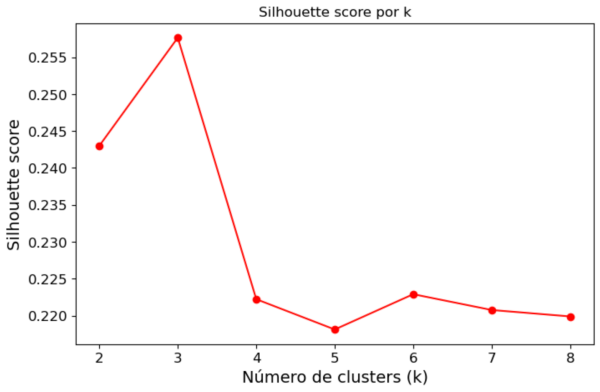
Dado que durante el análisis exploratorio y los intentos de clustering jerárquico se observó que las **variables categóricas** codificadas mediante *OneHotEncoder* tenían un peso mucho mayor en las distancias que las variables numéricas, así como CapitalGainLog, decidimos buscar una solución que redujese estos pesos.

Para abordar este desequilibrio y eliminar la redundancia entre variables correlacionadas, se decidió aplicar dos técnicas de reducción de dimensionalidad complementarias:

- PCA (Principal Component Analysis), se aplicó sobre las variables numéricas, ya escaladas. En este grupo se incluye CapitalGain, cuyo peso queremos reducir, por esta razón se usó como parámetro n_components = 3, pues tenemos 4 características y queremos reducir su dimensionalidad pero no demasiado.
- MCA (Multiple Correspondence Analysis), funciona sobre variables categóricas, ya sean strings o variables dummy. Se aplicó sobre las ya codificadas con OneHotEncoder. Buscamos comprimir ese gran número de variables binarias, interpretando nuestros datos, consideramos altamente posible que existiesen patrones de co-ocurrencia entre categorías (Male-Husband, Female-Wife, “Husband-Married”, etc.). Usamos n_components = 3 para equilibrar con PCA y no perder información excesiva, vimos que con más componentes los valores de silhoutte se mantenían bastante estables.

Los resultados de la evaluación de distintos valores de k sobre los datos reducidos son los siguientes:



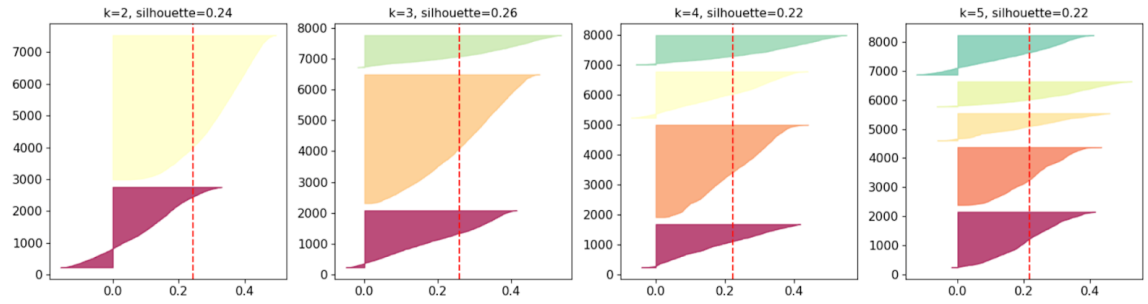


Vimos que la inercia no mostraba un "codo" relativamente marcado sin embargo vemos que el decrecimiento se disminuye a partir de los puntos donde k = 3 y k = 4, dando a entender que es posible que esta cantidad de clusters sea la óptima para representar el agrupamiento natural de los datos. No obstante, en el estudio de los coeficientes, se ven un claro pico en k = 3 con s = 0.258, lo cual apoyó hasta cierto grado el gráfico anterior.

Teniendo en cuenta ambos criterios, se seleccionó k = 3 como número óptimo de clusters, ya que proporciona una estructura interpretable y con separación razonable entre grupos.

Esta configuración se empleó en los análisis posteriores para interpretar las características de cada cluster y su correspondencia con las variables originales.

Observamos de forma gráfica los clusters resultantes para cada k con los diagramas de silhoutte, donde se muestra, para cada punto, su grado de pertenencia al cluster asignado en relación con los demás.
Vemos en general que los coeficientes de silhoutte se encuentran principalmente entre 0.1 y 0.4, lo que indica una separación moderada pero razonable entre los grupos. No obstante, apenas observamos scores negativos lo que sugiere que, aunque los clusters no están muy separados los unos de los otros, debido a que nuestros datos son mixtos, no hay un número importante de datos mal asignados, con la excepción de k = 2.

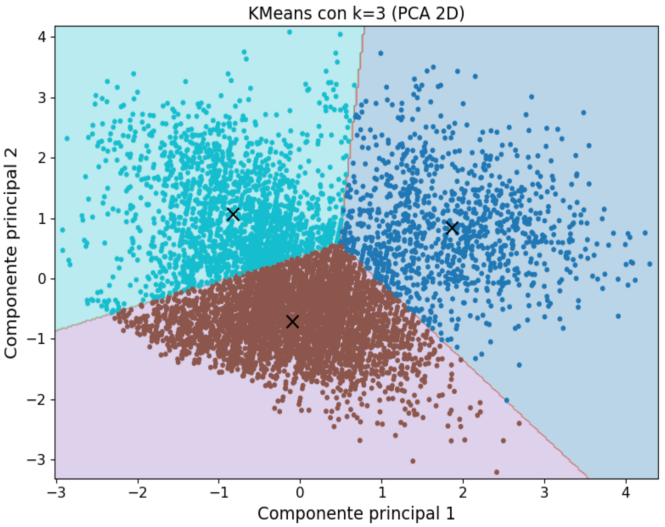


Centrándonos en el de k = 3, vemos que hay una mejora clara ante los demás valores. Las bandas son generalmente anchas y compactas, mostrando agrupaciones consistentes y puntos

relativamente bien cohesionados. Notamos que el segundo cluster es muy grande, contiene una gran porción del total de datos del conjunto, podríamos por ejemplo suponer que este se corresponde con las personas que no tienen ganancia de capital, sin embargo, después de la aplicación de PCA, podría no seguir cumpliéndose este patrón.

El valor medio global de 0.26, aunque no es muy alto, es típico en datos mixtos o de naturaleza socioeconómica, donde la separación perfecta es muy inhabitual. Consideramos que los coeficientes confirman que k = 3 es un clustering adecuado.

Aunque requiere una reducción de dimensionalidad de nuevo, podemos visualizar este clustering mediante una gráfica de fronteras de decisión.



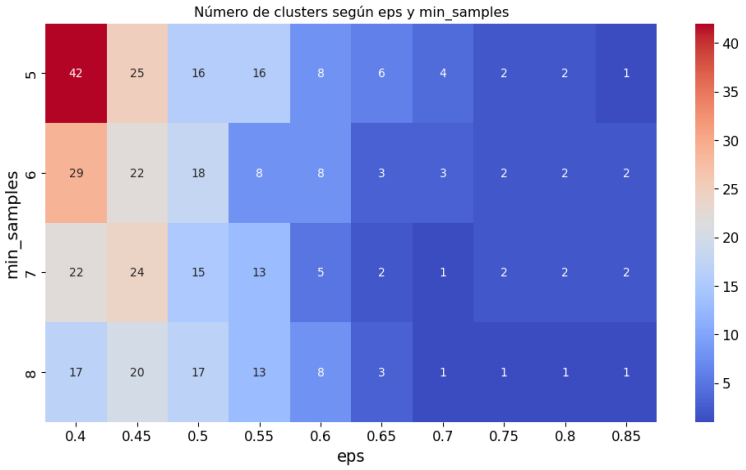
Parece que la división es proporcional, los centroides marcados están lo suficientemente separados los unos de los otros. Las fronteras son prácticamente lineales, esto no demuestra naturalidad, ya que es extraño que se dirán particiones tan exactas en datos reales, sin embargo debemos notar que esto es típico en KMeans, ya que se basa en distancias euclídeas.

Aunque las fronteras no capturan toda la complejidad del espacio original, debido a la reducción de dimensionalidad del PCA, la proyección evidencia que el modelo logra identificar tres agrupamientos consistentes.

3.3 DBSCAN

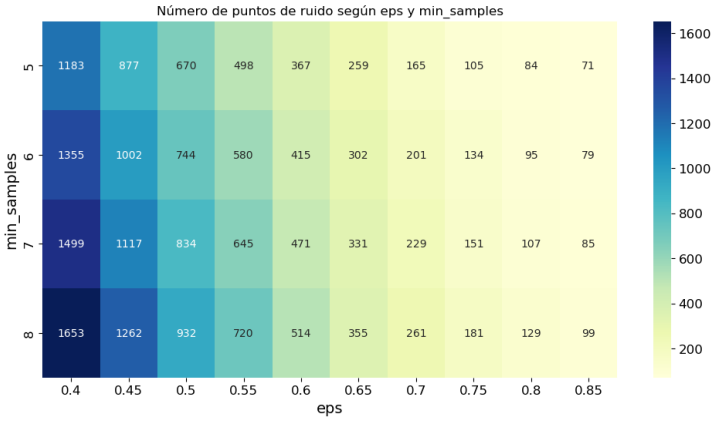
Para el estudio de este método se realizó un análisis del efecto de los parámetros eps y min_samples sobre los resultados del clustering. Una primera evaluación que nos sirvió como guía para hallar un buen rango de valores se llevó a cabo sobre valores de eps entre 0.2 y 2.1, y de min_samples entre 3 y 12. Para cada combinación, se calcularon tanto el número de clusters como el número de puntos de ruido, y se visualizaron mediante un heatmap.

Con el fin de simplificar la interpretación de los resultados, se generó un segundo heatmap seleccionando únicamente las combinaciones de parámetros que presentaban un equilibrio adecuado entre el número de clusters detectados y la cantidad de puntos de ruido. Esto permitió identificar de manera más clara los valores de eps y min_samples que generan un clustering significativo y representativo de la estructura de los datos.

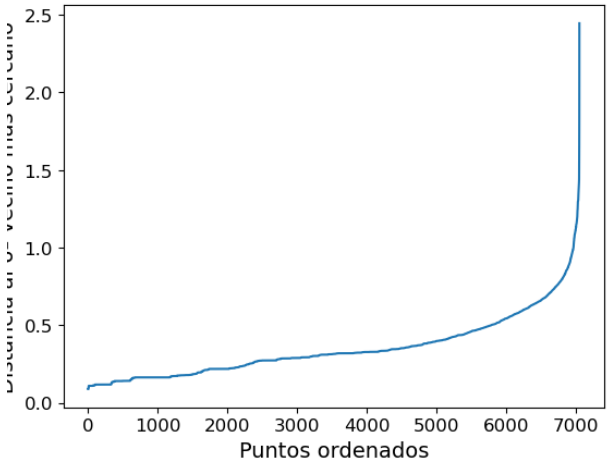


Para min_samples = 5, se observa la presencia de muchos microclusters a medida que aumenta eps, lo que indica que la agrupación está demasiado fragmentada. Al aumentarlo a 6 o 7, el número de clusters se estabiliza más rápidamente, entre 3 y 5, para valores de eps alrededor de 0.6 - 0.7. En cambio, con min_samples = 8, la estructura se vuelve demasiado rígida, colapsando rápidamente a un único cluster.

En cuanto al ruido, min_samples = 5 genera menos puntos de ruido (aproximadamente entre 500 y 800). Los valores min_samples = 6 y 7 producen más ruido (1000–1500 puntos), pero a cambio obtienen clusters más limpios y mejor definidos. Por su parte, min_samples = 8 genera un exceso de puntos de ruido (más de 1600), lo que sugiere que está filtrando en exceso y eliminando información relevante. Se eligió 7 como valor definitivo.



Para la evaluación de distintos valores de eps y su efecto en el clustering resultante, buscamos un candidato inicial. Visualizamos la gráfica de la distancia al 7º vecino más cercano, se eligió analizar la distancia al 7º vecino porque el parámetro min_samples se estableció en 7. Esta gráfica permite identificar el valor de eps adecuado para DBSCAN, mostrando la distancia a partir de la cual los puntos empiezan a considerarse aislados y, por tanto, ruido. El "codo" de la curva indica un rango de densidad óptimo para formar clusters, facilitando la selección de un eps que agrupe correctamente los puntos densos del dataset.



Basándonos en que el "codo" de la gráfica está entre las distancias 0.6 y 1.2, podemos estimar un valor de eps usando la distancia al 7º vecino correspondiente en el eje Y en ese rango.

Los resultados mostraron que los valores de eps más pequeños, como 0.6, 0.7 y 0.8, producen muchos clusters y un alto porcentaje de ruido, con valores negativos de Silhouette, lo que indica

que los clusters están solapados y poco estructurados. Por otro lado, valores mayores como 1.1 y 1.2 generan un único cluster, descartando la posibilidad de identificar subgrupos. En cambio, eps igual a 0.9 o 1.0 da lugar a pocos clusters con bajo porcentaje de ruido y valores positivos de Silhouette, lo que refleja agrupaciones más cohesionadas y bien definidas.

En conclusión, utilizando el Silhouette Score como criterio cuantitativo de calidad, eps = 1.0 se considera el valor más adecuado, ya que ofrece un buen equilibrio entre número de clusters, cohesión interna y mínima presencia de ruido.

El modelo detectó 2 clusters principales y un total de 80 puntos de ruido, lo que representa aproximadamente el 1.1% del dataset.

	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

Se observan diferencias claras entre los grupos:

- El Cluster 0 representa la mayoría del dataset, con edad promedio media, jornada laboral estándar y nivel educativo moderado.
- El Cluster 1 es muy pequeño, con individuos mayores y jornadas laborales reducidas, pero con altas ganancias de capital.
- Los puntos de ruido (-1) corresponden a casos atípicos o poco representativos, con edad y horas trabajadas más elevadas, y menor nivel educativo.

Para las variables categóricas (MaritalStatus, Relationship, Gender), se identificó la categoría más frecuente en cada cluster:

	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

Estas características permiten interpretar los clusters de manera cualitativa:

Cluster 0 agrupa a individuos casados, mayoritariamente hombres con rol de esposo.

Cluster 1 corresponde a personas viudas, todas mujeres y sin núcleo familiar.

Los puntos de ruido incluyen personas solteras, sin estructura familiar definida y con distribución de género equilibrada.

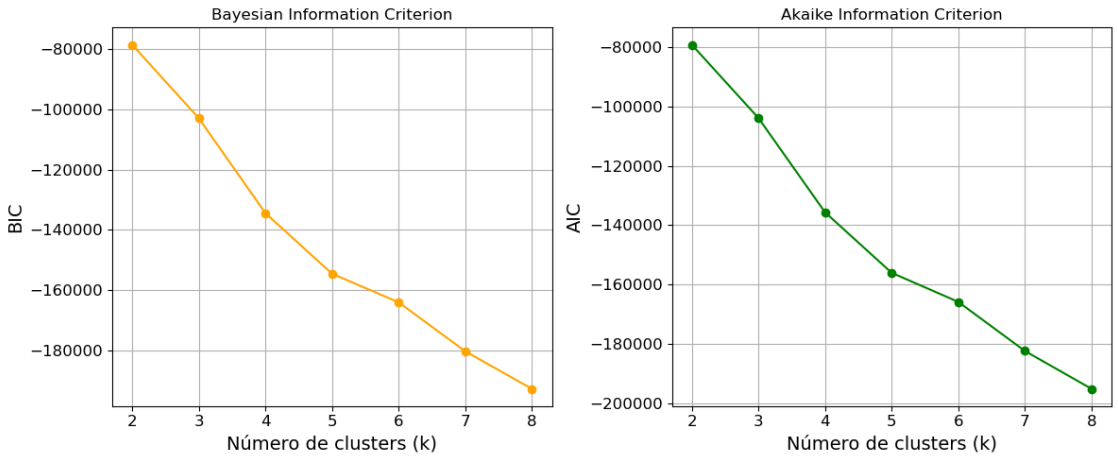
En conjunto, el clustering final con DBSCAN no permitió identificar la estructura principal del dataset si no que solo pudo separar un pequeño grupo de individuos significativamente diferentes.

El mal comportamiento de DBSCAN probablemente se deba a que este método busca detectar zonas de alta densidad separadas por zonas vacías, no obstante, dado que nuestros datos tienen alto solapamiento y valores continuos, lo más probable es que haya pocas zonas vacías y se distribuyan los datos como una gran nube.

1.5. Gaussian Mixture Models

Se utilizó Gaussian Mixture Models (GMM) para explorar la posible estructura de clusters en los datos, evaluando el número óptimo de componentes mediante tres métricas: Silhouette, BIC (Bayesian Information Criterion) y AIC (Akaike Information Criterion). Se probaron valores de k entre 2 y 8, entrenando cada modelo con 5 inicializaciones aleatorias para mayor robustez.

Los resultados obtenidos fueron los siguientes:



Se observan las siguientes tendencias como que el Silhouette es bajo en todos los casos, indicando que los clusters no están muy bien separados y que la estructura del conjunto de datos es difusa. Por otro lado, tanto BIC como AIC disminuyen a medida que aumenta el número de clusters, lo que es típico en GMM: modelos con más componentes ajustan mejor los datos, pero también aumentan la complejidad. Considerando un equilibrio entre complejidad y separación de clusters, k = 2 o 3 parecen las opciones más razonables, ya que los valores de Silhouette son relativamente más altos, mientras que BIC y AIC no disminuyen de forma dramática respecto al siguiente valor.

Las gráficas de BIC y AIC refuerzan esta interpretación, mostrando que los criterios de información disminuyen con k, pero la mejora se reduce a partir de k = 3, lo que sugiere que agregar más clusters aporta poco beneficio en términos de ajuste.

En resumen, GMM sugiere que el conjunto de datos podría dividirse en 2–3 clusters principales, aunque la separación entre ellos no es muy marcada, lo que coincide con las observaciones previas del análisis de PCA y DBSCAN.

Se calculó un resumen estadístico para cada cluster considerando variables numéricas de interés:

	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

Este resumen muestra diferencias claras entre los clusters:

- El Cluster 0 agrupa a personas más jóvenes con jornada laboral moderada y menor ganancia de capital.
- Los Clusters 1 y 2 agrupan a individuos mayores, diferenciándose principalmente en horas trabajadas y ganancias de capital, siendo el Cluster 2 el que tiene mayores valores de ambos indicadores.

Para las variables categóricas (MaritalStatus, Relationship, Gender) se identificó la categoría más frecuente en cada cluster y su proporción:

	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

Estos resultados permiten interpretar los clusters de manera cualitativa:

- Cluster 0 está compuesto principalmente por individuos solteros y sin núcleo familiar.
- Cluster 1 agrupa mayoritariamente a personas casadas y mujeres con rol de esposa.
  Cluster 2 contiene personas casadas y hombres con rol de esposo, reflejando un patrón muy definido de estructura familiar.

En conjunto, este análisis permite no solo identificar clusters en los datos numéricos, sino también caracterizarlos con variables demográficas y laborales, proporcionando una visión completa de la estructura del dataset.

4. Conclusión

Tras la realización de este proyecto, hemos extraído variadas conclusiones acerca de nuestro dato, nuestra metodología y nuestros resultados.

Primeramente, hemos aprendido la importancia del escalado de las variables y el preprocesamiento en general, comprobando que lograr un peso equitativo para todas las características es clave para obtener agrupaciones coherentes en vez de arbitrarias, evitando que cierta información domine sobre las distancias opacando a las demás. Se ha evidenciado la complejidad que requiere la toma de decisiones, teniendo muchas técnicas sobre las que elegir. Por ejemplo, elegimos aplicar los métodos de reducción de dimensionalidad PCA y MCA buscando minimizar el aplanamiento de las distancias causadas por las variables categóricas y

CapitalGain. Esto resultó fundamental para equilibrar la influencia de las distintas variables y asegurar que las distancias empleadas reflejaran similitudes reales entre los individuos. Sin embargo, reconocemos que esto conlleva una pérdida de información importante que a su vez afecta sobre el clustering. Por tanto como posible mejora podríamos explorar otras estrategias, como quizá la discretización de las variables numéricas en *bins* para que no sea tan drástica la diferencia entre los rangos continuos y los saltos binarios.

En cuanto a los métodos de clustering, pudimos observar las diferencias significativas entre los resultados obtenidos, viendo como muchos nos indican que el clustering óptimo era de 2 clusters mientras que otros apuntaban a tres. Sin embargo, podemos concluir que los datos deben tener buenos agrupamientos naturales rondando esas cifras. Vimos como el clustering jerárquico mostraba una jerarquía significativa de los datos, no arbitraria, sin embargo se veía demasiado dominada por CapitalGain, seguida de las variables categóricas. Por otro lado tanto KMeans como Gauss Mixture Method lograban hallar clusters coherentes, donde se veían divisiones entre género, edad, estado civil, ganancia de capital, etc. Un método que resultó bastante insuficiente fue DBSCAN, que encontró dos clusterings que si bien devuelven un buen coeficiente silhouette, uno de dichos grupos constaba de sólo 6 individuos.

Finalmente, otra posible mejora podría ser la exploración de distancias o funciones de disimilitud alternativas a la distancia euclídea, que fue constante a lo largo de nuestro proyecto.

References

[1] Wongoutong C. The impact of neglecting feature scaling in k-means clustering. PLoS One. 2024 Dec 6;19(12):e0310839. doi: 10.1371/journal.pone.0310839. PMID: 39642177; PMCID: PMC11623793.

[2] Fuente: Diapositivas “preprocessing” del Profesor Esteban García Cuesta, Profesorado (Coordinador), GCDIA – Aprendizaje Automático I, 2025.

[3] Fuente: Diapositivas “Unit 2: Clustering” del Profesor Esteban García Cuesta, Profesorado (Coordinador), GCDIA – Aprendizaje Automático I, 2025.