Práctica 1: Clustering y Preprocesado

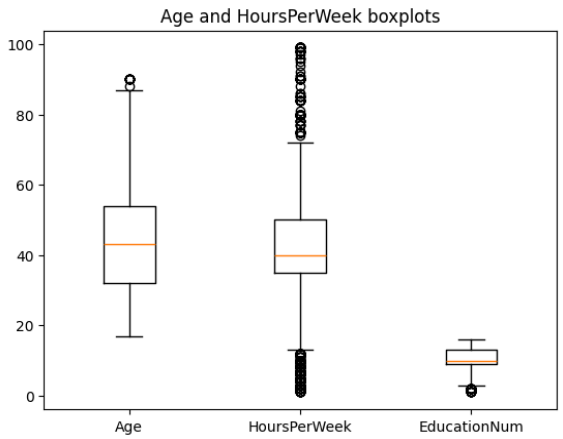
Autores: Joaquín Negrete, Pablo Suárez
Clase: 3º CDIA
Dataset: data-even.csv

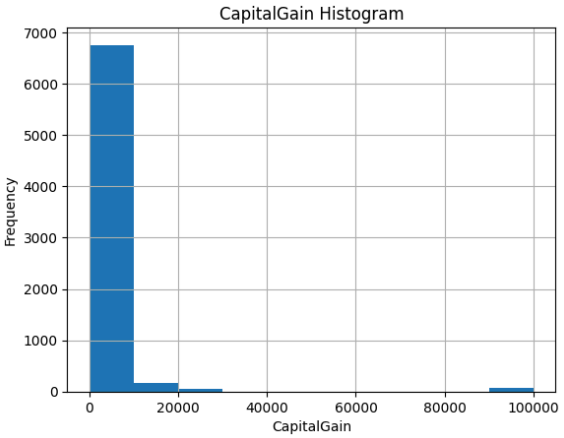
Preprocessing

Debemos hacer un análisis básico de la estructura de nuestros datos:

• age – age in years. (numeric, int)
• education – highest level of education attained. (string)
• education_num – years of education (numeric version of education) (numerical, int)
• marital_status – marital status (e.g., married, single, divorced). (string)
• relationship – family role within the household (e.g., husband, not-in-family). (string)
• gender – gender. (categorical,binary)
• capital_gain – income from capital gains. (numeric, int)
• hours_per_week – hours worked per week (numeric, int)

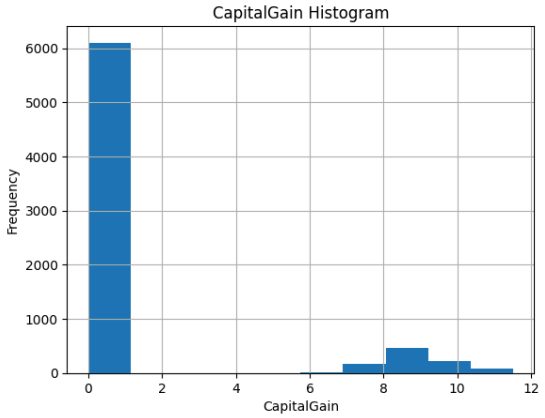
Primero comprobamos que no había ningún valor nulo, lo cual nos facilita el trabajo al no tener que usar ningún Imputer, también eliminamos algunas filas que estaban duplicadas.
Después analizamos la distribución de las variables numéricas (separamos CapitalGain de las demás por su rango tan alto de valores):





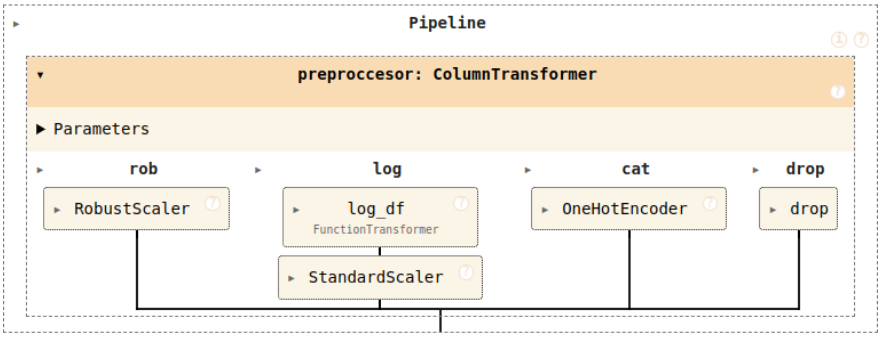
Podemos observar que la variable HoursPerWeek contiene un gran número de valores atípicos. Además, la distribución de CapitalGain es muy curiosa, un 86.49% de las filas contienen 0 's, pero tiene un rango muy alto de valores, siendo el máximo valor de 99999. Estas dos variables nos pueden causar muchos problemas, especialmente la de CapitalGain.

Para solucionar esto vamos a depender de nuestro escalado. Cómo hemos visto que, en general, nuestros datos tienen una gran variabilidad, hemos considerado que el mejor escalador que podríamos usar para las variables numéricas es RobustScaler, pues aunque StandardScaler tiende a funcionar bien en clustering, este es sensible a datos atípicos, y MinMax lo es aún más.



Sin embargo, no consideramos que esto fuese suficiente. Para la variable CapitalGain queríamos aplicar una modificación previa, aplicar la función log(1 +x), esta transformación mantendrá los valores bajos atrayendo los valores más altos, reduciendo así enormemente la escala de su separación.
Además nos dimos cuenta que, al tener tantos 0 ‘s, no podemos aplicar RobustScaler, pues su IQR y su mediana son iguales a 0. En su lugar aplicamos

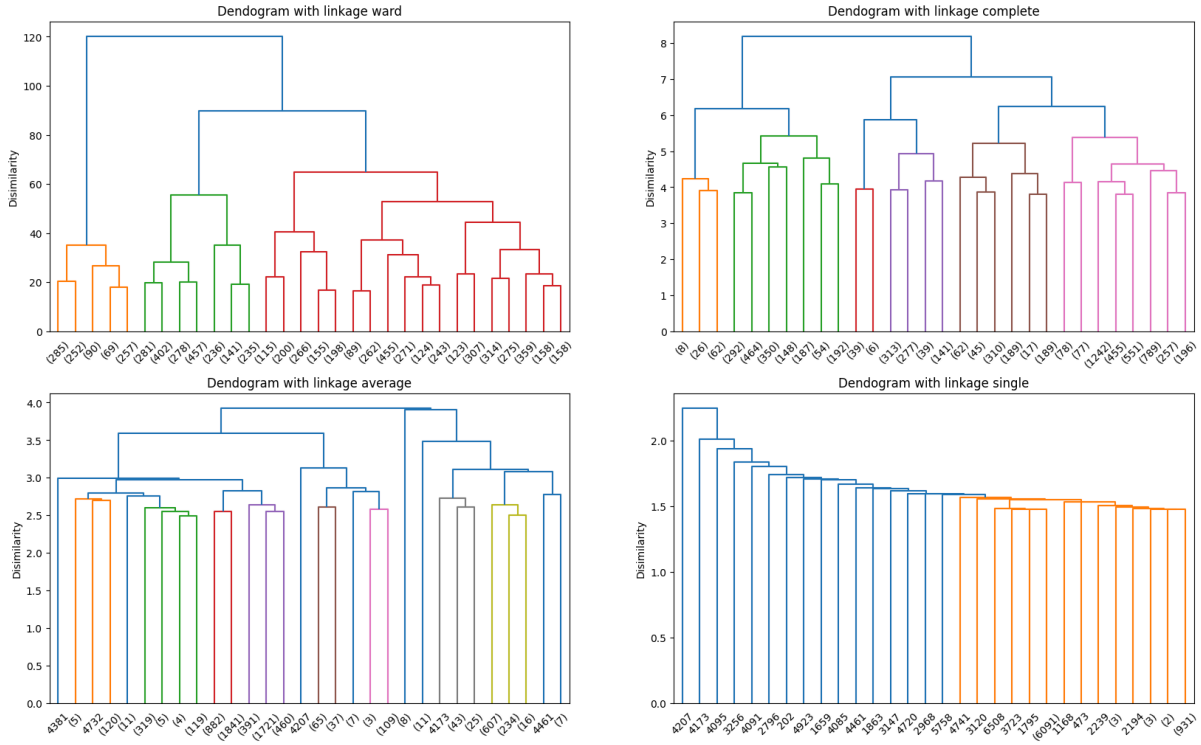
StandardScaler, pues pensamos que este, al centrar en 0, conjuntaría bien con las demás variables (a las que aplicamos RobustScaler).
Las variables MaritalStatus, Relationship y Gender son categóricas pero no son ordinales, por lo que la decisión es sencilla, usaremos OneHotEncoder. La variable Education es innecesaria teniendo EducationNum, por lo que la pasamos por drop.
Nuestro filtro de preprocesado queda entonces cómo:



Ahora que nuestros datos están preprocesados, probaremos a formar clusters con distintas técnicas:

Clustering Jerárquico

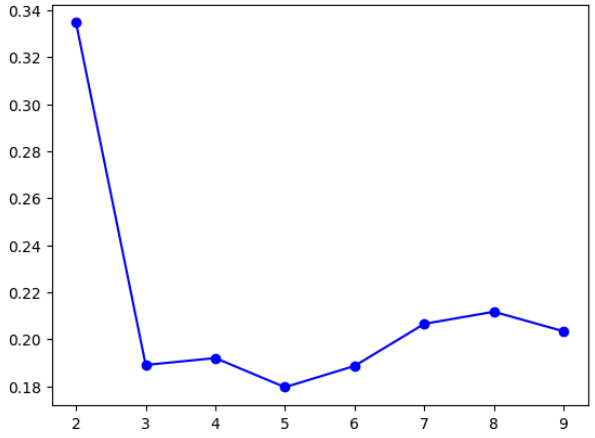
A la hora de hacer clustering jerárquico tenemos que escoger el número de clusters y el método de linkage, lo haremos principalmente mirando el dendrograma. Los resultados fueron los siguientes:



Podemos observar que el dendrograma más estructurado es, con diferencia, el que utiliza el método de ward. Además nos crea una muy buena separación entre clusters en los niveles más altos, que es lo que más nos importa. Por esta razón escogemos ward cómo el mejor linkage. Es fácil observar en este dendrograma las propiedades de los distintos linkages. En el single hay un claro caso de

encadenamiento, donde los clusters iniciales están muy dispersos y luego se unen todos al final. El completo no funciona nada mal, está creando clusters bastante regulares y compactos, también hubiera sido una buena elección. Average parece funcionar algo irregularmente en este caso.

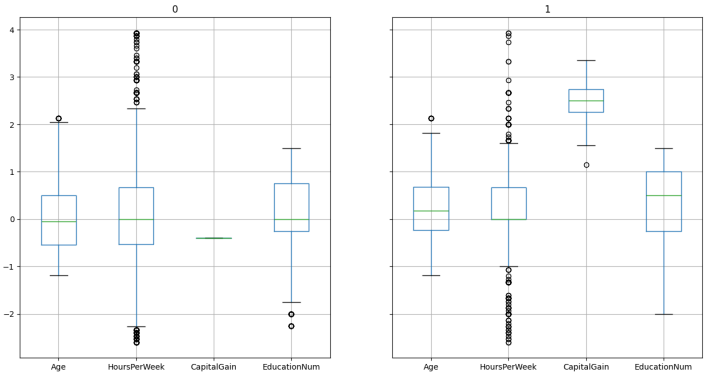
Ahora tenemos que escoger la mejor K, la cual parece ser 2 o 3. Tenemos 2 métodos, mediante el silhouette scores o interpretando cada resultado (mejor pues lo importante es al final sacar información de los clusters).



Simplemente observando el gráfico de coeficientes de silhouette parece que el mejor es 2 con mucha diferencia, sin embargo sospechamos que si escogemos 2 clusters, estos dos estarán simplemente separados por CapitalGain por su gran variabilidad, y esto no nos proporciona mucha información (como norma general evitaremos clusters que sólo nos haga una separación con el CapitalGain pues creemos que no nos aporta información nueva ni ninguna tendencia nueva).

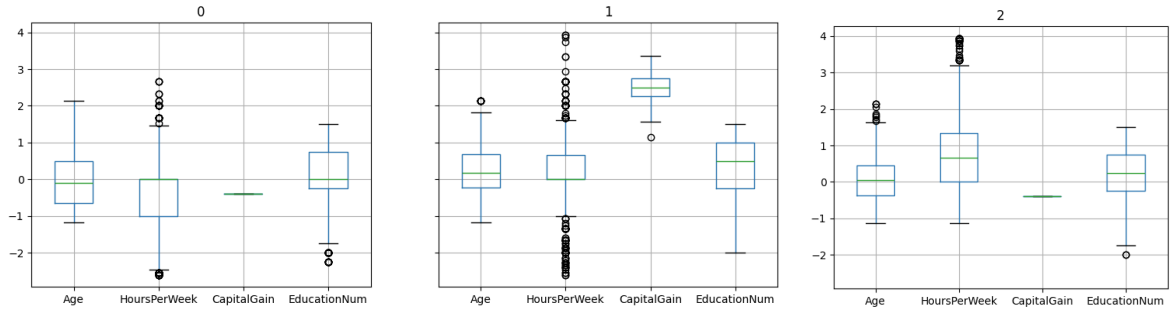
Comprobemos si esto es cierto.
Creamos ahora los clusters con K=2 y K=3 e interpretamos los resultados:

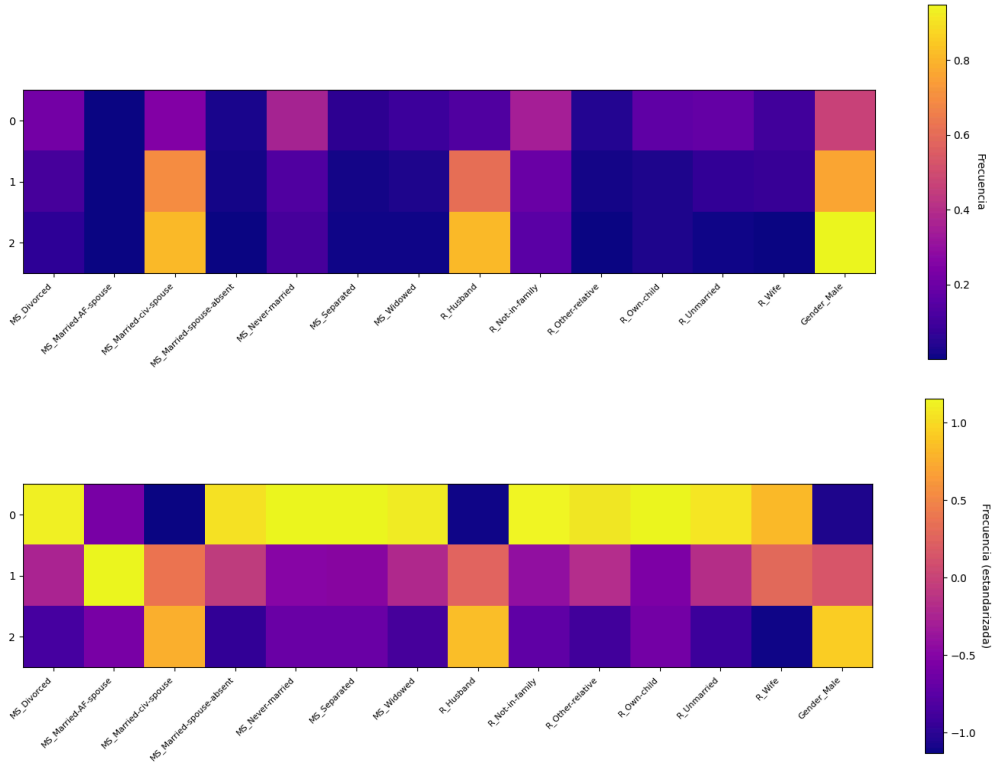
K=2:



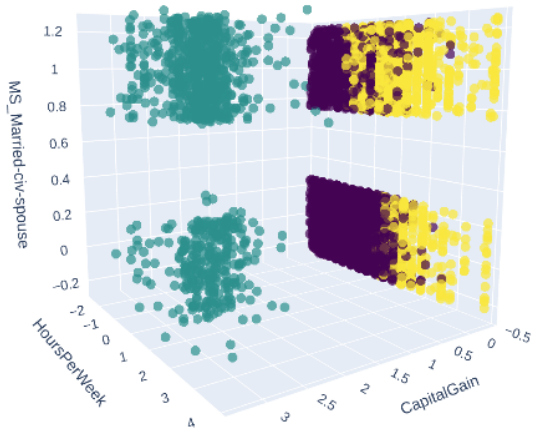
Podemos ver que, cómo sospechábamos, a la hora de crear 2 clusters, estos 2 se separan principalmente usando CapitalGain. Esto se aprecia gracias a la diferencia de medias. Este resultado es normal por la gran separación de la variable en dos grupos.

K=3:





Aquí podemos verdaderamente ver ciertos patrones en los datos, vemos que se separa en 3: **0)** Bajo capital y compuesto por una mezcla de hombres y mujeres (más mujeres) que no están casados. **1)** compuesto por la gente con gran ganancia capital, relativamente equilibrado entre hombres y mujeres y están muchos casados (muchos casados armed-forces) y **2)** principalmente hombres con bajo capital que están casados. Podemos ver que ha separado el grupo que no tenía capital en 2 grupos, hombres casados y mujeres (principalmente) sin casar (nunca casadas, viudas…) . También vemos que el grupo 0 trabaja menos que el grupo 2. Esto nos indica que las variables más importantes a la hora de separarse pueden ser: Si están o no casados, su género, su ganancia capital y sus horas trabajadas. Para mostrarlo:

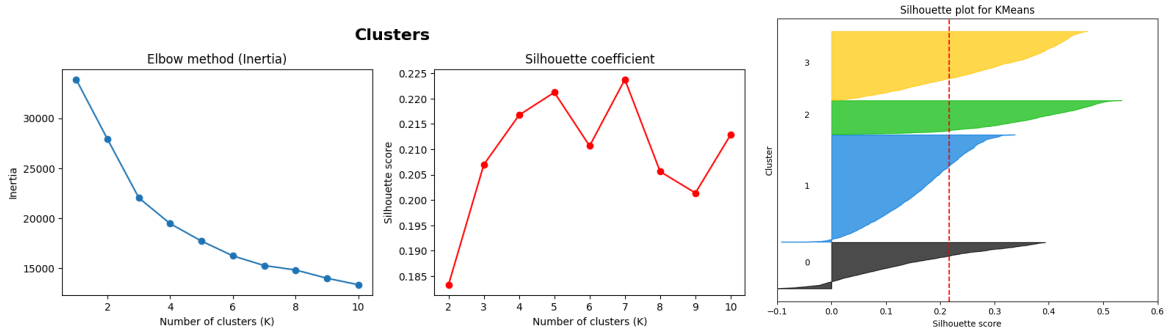


Nótese que se ha añadido una dispersión a las variables binarias para poder apreciar los puntos, pues sino parece haber un solo punto.

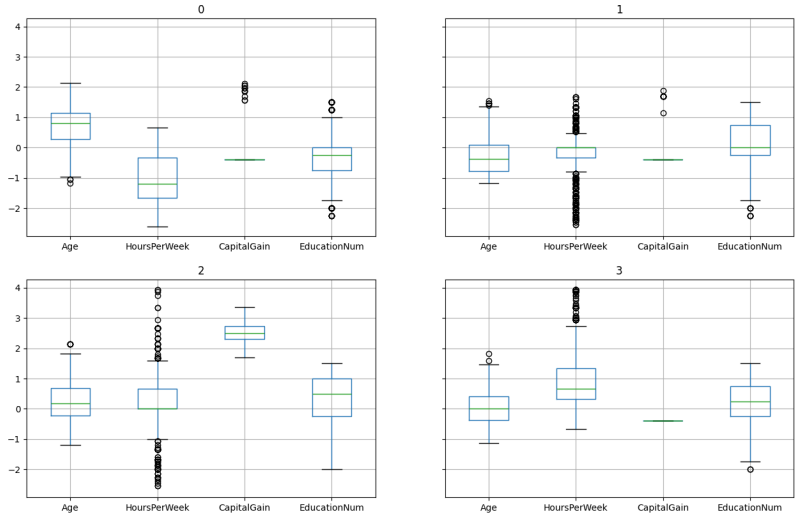
Consideramos que K=3 y “ward” es mejor por tener una separación más significativa.

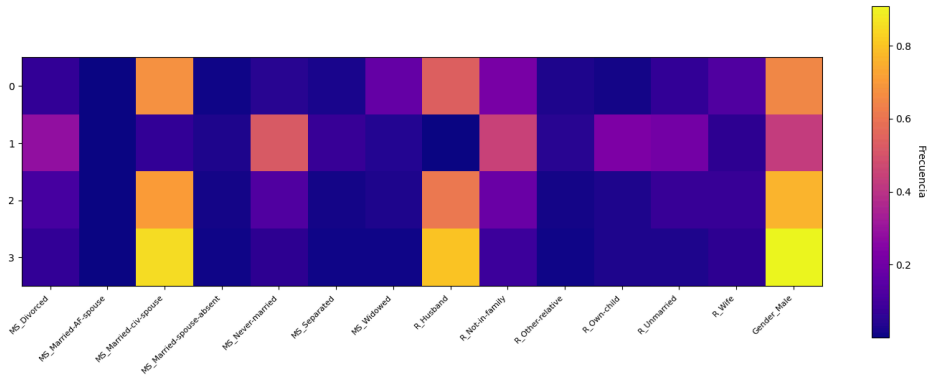
Clustering Particional

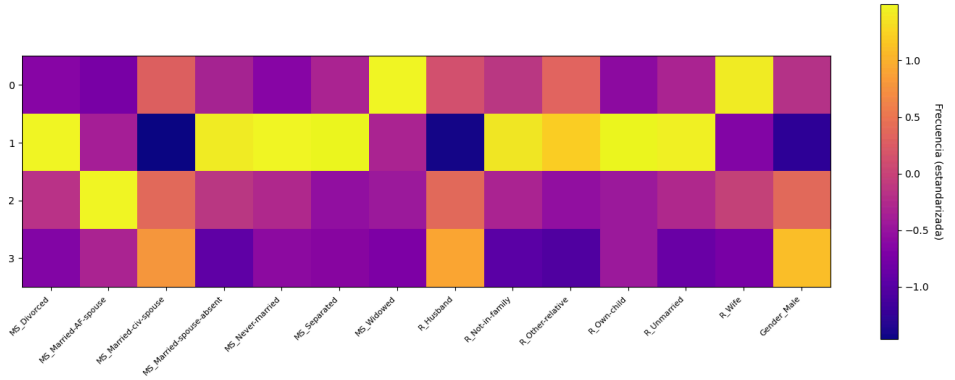
Ahora probaremos a usar K-means para nuestros clusters. De nuevo tenemos que escoger la K, lo haremos viendo el silhouette score y usando el método del codo:



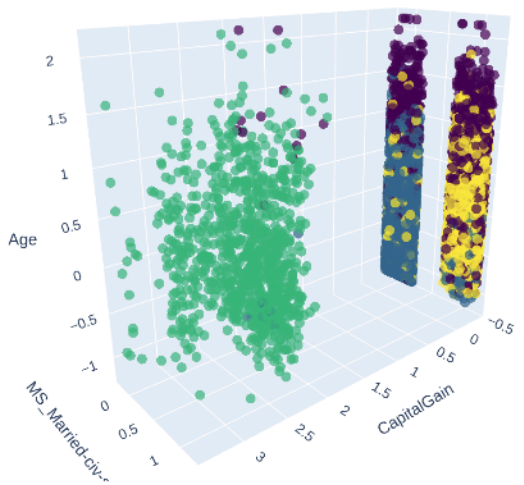
Analizando las dos gráficas podemos observar que hay dos valores para la K muy buenos, el 3, donde se forma el primer codo, y el 4, en el que el silhouette score aumenta sustancialmente. Ambos valores son aceptables, pero en este caso decidimos escoger K = 4.







- Cluster 0: tiene una alta proporción de mujeres viudas (MS_Widowed) y mujeres casadas,(R_Wife), y una baja proporción en el resto. Tiene un capital bajo y muy pocas horas de trabajo en comparación con el resto. También vemos que el grupo 0 es de edad más avanzada (tiene sentido entonces que haya mayor proporción de viudas y menos trabajo, puede hacer referencia a gente anciana o de edad avanzada).
- Cluster 1: tiene una alta proporción de gente soltera (divorced, married-spouse-absent, never-married, separated, not-in-family, own-child, and unmarried), y baja proporción de casados y en general también de hombres. Vemos que son de edad temprana, tienen una cantidad promedio/baja de horas trabajadas y una educación ligeramente por debajo de la media. (esto puede hacer referencia a la gente joven, lo que justifica los pocos matrimonios y la educación no finalizada de algunos)
- Cluster 2: tiene una alta proporción de gente casada por armed-forces, proporción media de married-civilian-spouse, husband, y male, baja proporción en el resto. Parece estar dominado por hombres, donde casi todos están casados. Este tiene una ganancia de capital muy alta, tienen un buen nivel de educación y un rango de edades muy variado y centrado ligeramente por encima de la media. El rango de horas trabajadas también tiene un rango muy alto. Parece que este grupo está determinado exclusivamente por el hecho de tener alguna ganancia capital, lo cual parece ser más común en hombres.
- Cluster 3: está compuesto casi en su totalidad de hombres casados de edad media/baja. Parece ser parecido al cluster 1, con la diferencia de estar dominado por hombres y de estar casados. También tienen una cantidad considerablemente alta de horas trabajadas. Quizás este grupo puede hacer referencia a gente de clase obrera.



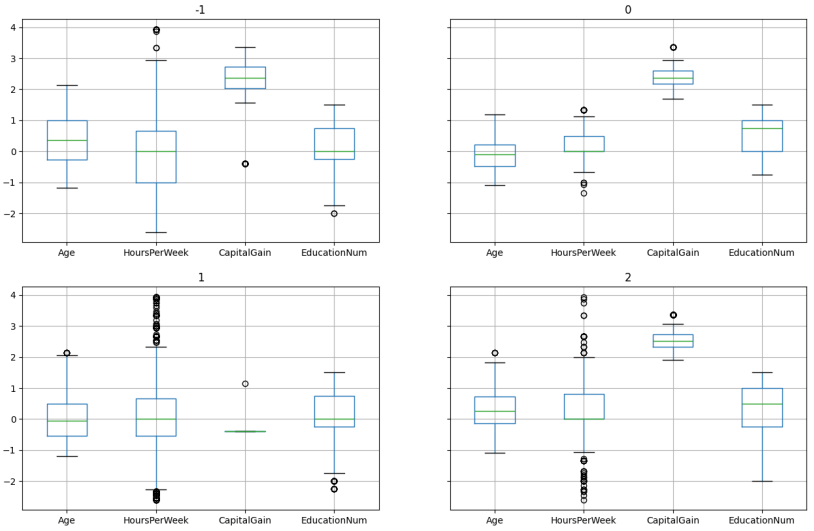
Cluster 0: morado
Cluster 1: azul
Cluster 2: verde
Cluster 3: amarillo

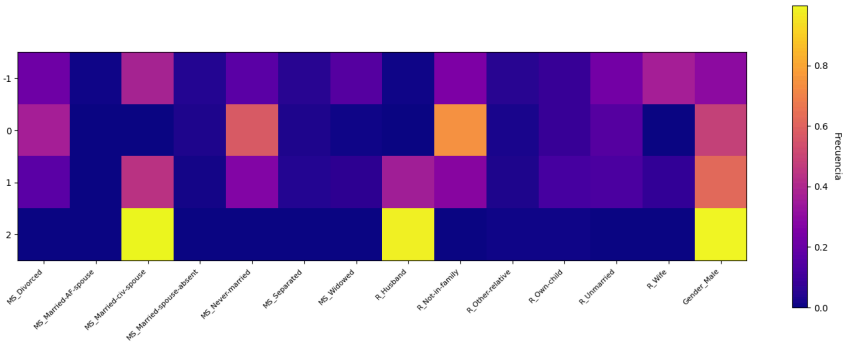
DBSCAN

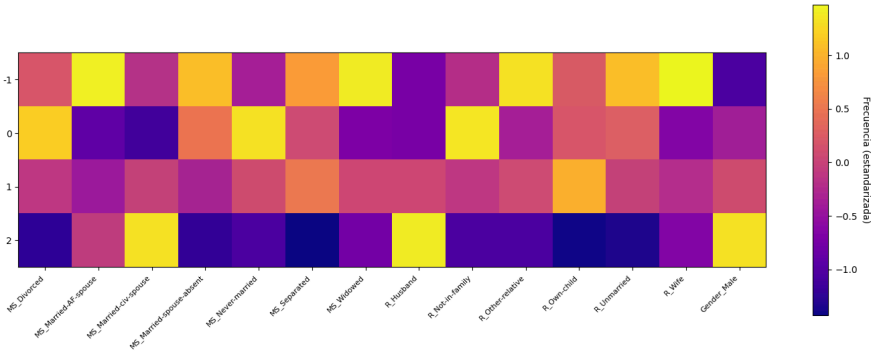
Ahora probaremos con el método de dbscan. Lo bueno es que ahora no tenemos que escoger la cantidad de clusters sino que lo hará el algoritmo automáticamente, lo malo es que tenemos que escoger ahora 2 parámetros, el epsilon y min_samples. Para ello hemos hecho for loops anidados probando con cada uno y calculando silhouette scores, después nos guardamos todos en un dataframe y vemos cual nos da el mejor valor (simulando Gridsearch). Los 10 mejores resultados fueron los siguientes:

eps	min_samples	n_clusters	silhouette_avg
1.65	70	3	0.270103
1.45	20	4	0.264489
1.85	120	3	0.261348
1.45	120	4	0.174309
1.45	170	4	0.172186
1.45	370	3	0.166416
1.45	220	4	0.166162
1.25	120	7	0.165121
1.45	270	4	0.159979
1.25	170	7	0.153623

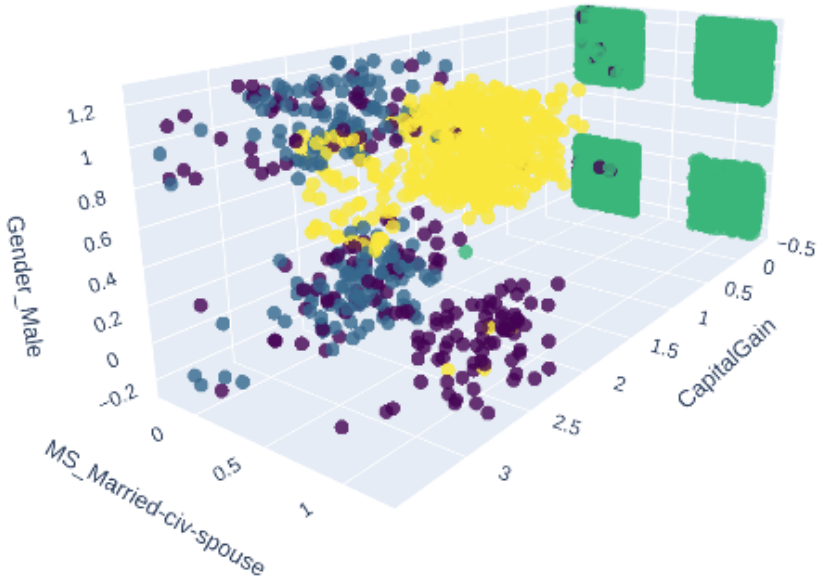
Vemos que los mejores resultados son epsilon = 1.65 y min_samples = 70. También vemos que la cantidad de componentes en el mejor resultado es de 3 componentes. Esto es bueno por ser consistente con nuestros anteriores intentos. Ahora creemos el modelo con esos parámetros y vamos a interpretar si tienen, o no, sentido.







Podemos ver que por primera vez se han creado 2 clusters que contienen a gente con ganancia capital alta:
El cluster 1 es el único con capital bajo, todos los demás valores son promedio, mezcla homogénea de hombres y mujeres casados y solteros… También las horas trabajadas, edad y educación están en la media.
Los clusters 0 y 2 tienen ganancia de capital alta y un nivel de educación superior. Sin embargo estos dos se diferencian por familia, el grupo 0 está formado por hombres y mujeres que no se han casado nunca y que no tienen familia, mientras que el grupo 2 está formado principalmente por hombres casados.
El -1, el cual corresponde con los valores atípicos. Estos no tienen porque seguir ninguna tendencia pues son todos los que no estén dentro de los otros clusters, estén donde estén. Sin embargo, podemos observar que se trata principalmente de mujeres, de las cuales bastantes están casadas.



Cluster 0: azul
Cluster 1: verde
Cluster 2: amarillo
Valores atípicos: morado

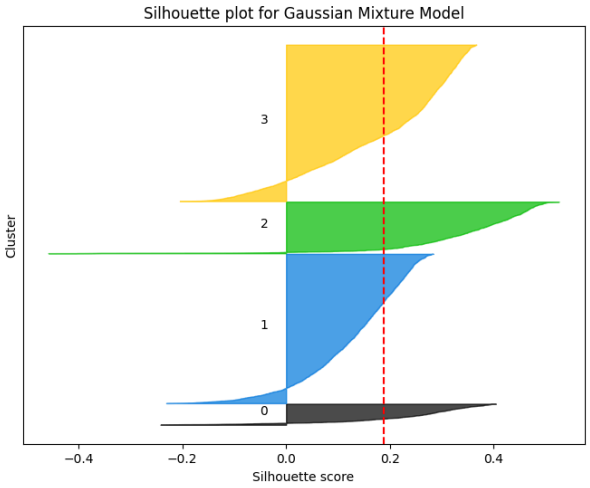
En este plot los valores atípicos se solapan mucho con los valores del cluster 0 porque estamos proyectando un espacio de altas dimensiones en solo 3 dimensiones, pero vemos que no hay casi soleamiento en los clusters 0, 1 y 2. El hecho de poder conseguir esto en un plot de 3 dimensiones es demostración de que gran parte de la variabilidad del modelo se explica con unas pocas variables (las cuales parecen repetirse en varios de los algoritmos que usamos).

Gaussian Mixtures

El último modelo que queremos probar antes de seleccionar un modelo final es el de mezclas gausianas. Para este modelo debemos seleccionar 2 parámetros distintos: el número de componentes y el tipo de covarianza. Además tenemos 2 métricas distintas que podemos usar para calcular los mejores parámetros: AIC y BIC. Para calcular el mejor haremos cómo antes, pero ahora haremos dos GridSearch distintos, uno usando AIC y el otro con BIC, después veremos los resultados. Los resultados de los dos GridSearch's son:

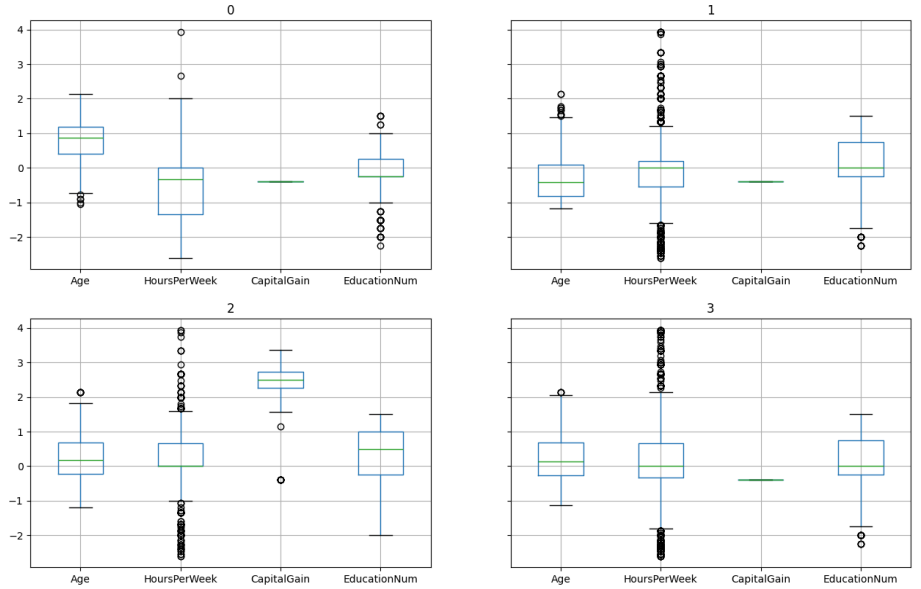
Best GMM parameters (BIC): {'covariance_type': 'full', 'n_components': 4}
Best GMM parameters (AIC): {'covariance_type': 'full', 'n_components': 4}

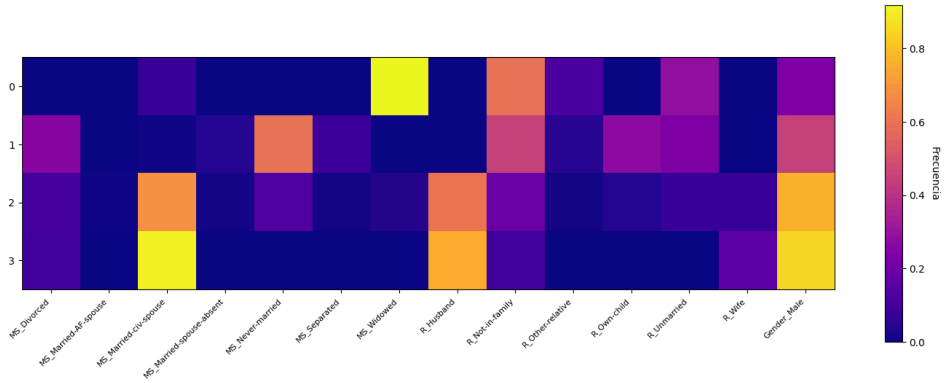
Podemos ver que ambos coinciden en los parámetros óptimos: tipo de covarianza = full y número de componentes = 4, lo que nos da más seguridad de que son los mejores parámetros.

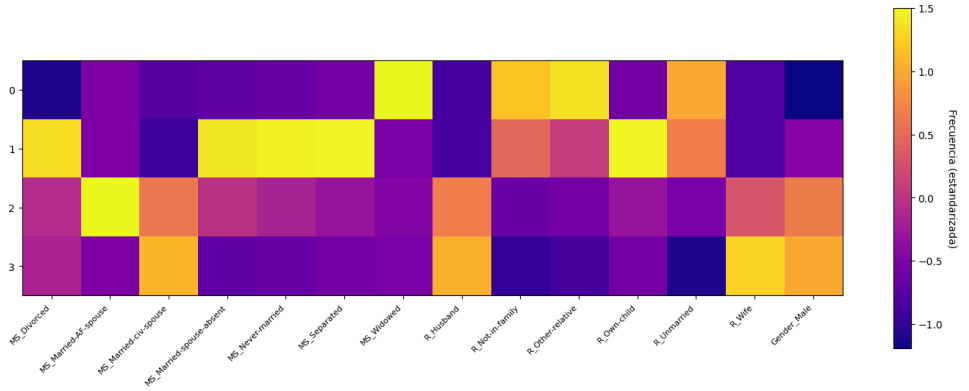


A pesar de ser 4 el valor óptimo de componentes, vemos que tampoco tiene un valor de silhouette especialmente bueno, esto también ha sido así en los modelos anteriores y puede deberse a que es un dataset difícil de separar.

Ahora, cómo en los anteriores modelos, pasemos a interpretar para ver qué tan significativos son los clusters:

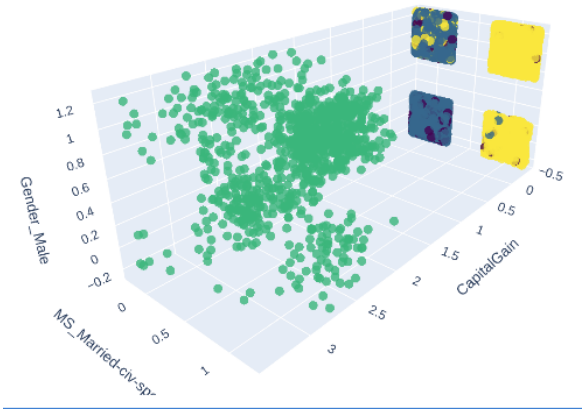






Vemos que se forman 4 clusters:

- Cluster 0: baja ganancia de capital, principalmente parece que se caracteriza o se separa por tener baja cantidad de horas de trabajo y ser de edad avanzada. Se trata principalmente de mujeres sin familia.
- Cluster 1: Este, aunque también tiene ganancia de capital baja parece ser algo contraria, pues está formada por gente joven. Es una mezcla homogénea de hombres y mujeres que no están casados.
- Cluster 2: Aquí tenemos al grupo que sí tiene ganancia capital. Como siempre, hay mayor cantidad de hombres y además están casados.
- Cluster 3: Formado principalmente por hombres casados.



Cluster 0: morado
Cluster 1: azul
Cluster 2: verde
Cluster 3: amarillo

Podemos ver que aunque usemos distintos métodos de cluster, gran parte de los clusters que se forman se repiten. Además muchos de ellos dependen de una cantidad limitada de variables. Algunas de las variables con mayor importancia pueden ser CapitalGain, HoursPerWeek, Married-civ-spouse. Podemos intentar comprobar esta intuición con PCA.

PCA

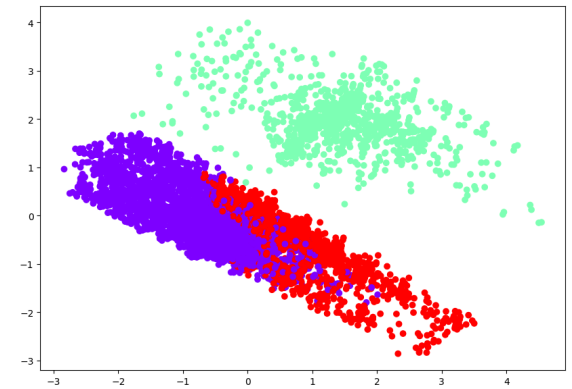
Para comprobar si es cierto comprobamos los componentes principales del dataset con PCA:

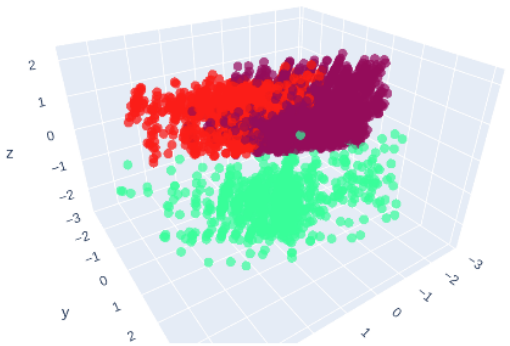
	PC1	PC2	PC3
Age	0.047177	0.249875	0.474095
HoursPerWeek	0.802514	-0.546822	-0.065090
EducationNum	0.169330	0.044252	-0.196721
CapitalGain	0.454009	0.762975	-0.413862
MS_Divorced	-0.023423	-0.044270	-0.094406
MS_Married-AF-spouse	0.000037	0.000150	-0.000578
MS_Married-civ-spouse	0.182203	0.140043	0.413605
MS_Married-spouse-absent	-0.003074	-0.002707	-0.007090
MS_Never-married	-0.107495	-0.097981	-0.307359
MS_Separated	-0.012120	-0.011874	-0.015828
MS_Widowed	-0.036127	0.016639	0.011656
R_Husband	0.197423	0.120458	0.409593
R_Not-in-family	-0.069943	-0.066851	-0.215460
R_Other-relative	-0.012750	-0.003364	-0.012192
R_Own-child	-0.064565	-0.043388	-0.118445
R_Unmarried	-0.035564	-0.026221	-0.067691
R_Wife	-0.014601	0.019366	0.004195
Gender_Male	0.150031	0.030435	0.230564

Podemos ver que, en efecto, el primer y segundo componente están bastante determinados por HoursPerWeek. El segundo también está muy determinado por CapitalGain, que es el número máximo en todo el PCA. En el segundo influencian mucho la edad, el CapitalGain, el Married-civ-spouse y si es, o no, viuda. Se comprueban nuestras sospechas de que esas variables influencian mucho a nuestros clusters (aunque se nos escapó la variable de viudas).

Modelo Final

Cómo ya hemos mencionado, los clusters que hemos ido consiguiendo tienen bastante sentido y encuentran patrones bastante parecidos. El único con algo de diferencia es el de dbscan. Los dos que parecían separar mejor los datos son el jerárquico y el de mezclas gaussianas. Cómo modelo final hemos decidido escoger el modelo jerárquico. Esto es pues ha hallado un modelo sencillo, con sentido y bien definido, usando las principales variables más importantes, HoursPerWeek, CapitalGain y Married-civ-spouse. PCA coloreado con los labels del jerárquico:





Resultado y conclusiones

Nuestro resultado concluye con 3 clusters formados usando clustering jerárquico y usando ward cómo linkage. Hemos conseguido un modelo fácil de explicar pero con gran significatividad. Se comprueba que son buenos clusters pues tienen sentido y salen repetidos en diversos algoritmos. Además, indirectamente hemos conseguido analizar cuales son las variables que más explican la variabilidad de nuestro modelo. Estamos contentos de que creemos que podemos decir que tenemos una buena comprensión del modelo y de ciertas tendencias en él. Hemos aprendido muchas técnicas a la hora de cuestionarnos cómo interpretar el modelo. También el trabajo nos hizo plantearnos la duda de cómo distinguir entre un mejor o peor clustering.