PIB DE PAÍSES Y SU FUERZA LABORAL A TRAVÉS DE LOS AÑOS

POR:
Nicolas Eduardo Perez
Norbey Garcia Arbelaez

MATERIA:
Introducción a la Inteligencia Artificial

PROFESOR:
Raul Ramos Pollan

UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
MEDELLÍN 2021

INTRODUCCIÓN

A través de la inteligencia artificial y gracias a sus avances, al dia de hoy puede ser utilizada en diversos campos y aplicaciones, una de estas es la generación de modelos de predicción, los cuales nos ayudan a estimar de manera aproximada o incluso de manera exacta, cálculos o respuestas que queremos obtener, todo esto mediante unos datasets, que nos proporcionan miles de datos recopilados a los cuales se le puede hacer un análisis estadístico y generar una predicción de datos y a su vez esos datos obtenidos pueden ser parte de un nuevo dataset actualizado.

1. PLANTEAMIENTO DEL PROBLEMA

Es importante conocer la fuerza laboral tanto de mujeres como de hombres en diferentes países y de cómo se relaciona con el producto interno bruto, esto para entender un poco las brechas en la participación laboral entre hombres y mujeres que hay en el mundo, es por eso que intentaremos desarrollar un modelo que nos permitiera entender muy bien esta problemática y poder cuantificar, para que los países puedan implementar políticas globales para cerrar dicha brecha.

Variable objetivo
Agrupar los datos de varios países basados en los PIB similares para luego dar lectura en estos grupos similares de el cambio de la fuerza laboral en comparación de los géneros a lo largo de los años.

1.1 DATASET

Para el dataset se importaron datos del PIB y que contenían datos de la calidad de vida de varios países, los cuales se importaron de la página de World Bank. A continuación se muestran los links y nombres de los archivos:

Datos de las mujeres:
https://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS
Datos de hombres:
https://data.worldbank.org/indicator/SL.TLF.ACTI.MA.ZS
PIB:
https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

La importación que contiene los datos se hizo utilizando el código mostrado a continuación.

```
[2]  auth.authenticate_user()
     gauth = GoogleAuth()
     gauth.credentials = GoogleCredentials.get_application_default()
     drive = GoogleDrive(gauth)

     file_id = '1h7BE5hPVF_dCZiqelgPMdpuyujbZblbg'
     download = drive.CreateFile({'id':file_id})
     download.GetContentFile('Male.csv')

     file_id = '16uNZicJTSPGNSqrVgD7I6nkN-kdV3npf'
     download = drive.CreateFile({'id':file_id})
     download.GetContentFile('Female.csv')

     file_id = '1USrX33Uk0rCEOZrKREf-rTFUrOZh0mao'
     download = drive.CreateFile({'id':file_id})
     download.GetContentFile('GDP.csv')
```

Imagen 1: Importación del PIB global

También se importaron datos de la fuerza laboral masculina y femenina, a la que se le hizo una limpieza del conjunto de datos, ya que, habían algunos datos faltantes/defectuosos y columnas innecesarias en nuestro análisis que se limpiaron. Además para evitar la redundancia decidimos trabajar con menos datos anuales, para lo cual elegimos los años de 1990, 2010 y 2018 para el análisis. El código se muestra a continuación.

```
male = pd.read_csv('Male.csv', delimiter=',', on_bad_lines='skip', header=1)
unnamed_cols_male = male.columns.str.contains('Unnamed')
male_clean = male.drop(male[male.columns[unnamed_cols_male]], axis=1)
male_clean = male_clean.dropna()
male_clean.drop([192,193,194,195,196,198,199,200,201,202,203,204,205,206,207],
                axis=0, inplace=True)
male_clean.replace("..", np.nan, inplace=True)
male_clean.replace(" ", np.NaN, inplace=True)
male_clean.drop(["1995","2000","2005","2011","2012","2013","2014",
                 "2015","2016","2017"], axis=1, inplace=True)

female = pd.read_csv('Female.csv', delimiter=',', on_bad_lines='skip', header=1)
unnamed_cols_female = female.columns.str.contains('Unnamed')
female_clean = female.drop(female[female.columns[unnamed_cols_female]], axis=1)
female_clean = female_clean.dropna()
female_clean.drop([192,193,194,195,196,198,199,200,201,202,203,204,205,206,207],
                  axis=0, inplace=True)
female_clean.replace("..", np.NaN, inplace=True)
female_clean.replace(" ", np.NaN, inplace=True)
female_clean.drop(["1995","2000","2005","2011","2012","2013","2014",
                   "2015","2016","2017"], axis=1, inplace=True)
```

Imagen 2: Importación de datos de fuerza laboral.

También limpiamos el .CSV con los datos del PIB, para lo cual eliminamos las columnas que no tenían nombre y eliminamos columnas con datos vacíos. El código es mostrado a continuación:

```
[4]  GDP = pd.read_csv('GDP.csv', delimiter=',', on_bad_lines='skip', header=0,
                       names=["Pais","Continente","Pib"])
     unnamed_cols_GDP = GDP.columns.str.contains('Unnamed')
     GDP_clean = GDP.drop(GDP[GDP.columns[unnamed_cols_GDP]], axis=1)
     GDP_clean = GDP_clean.dropna()
```

Imagen 3: Limpieza y depuración de datos del GDP.

Luego convertimos los datos anuales y de IDH de datos de la fuerza laboral de dataframes a float para facilitar el trabajo, también llenamos los datos que faltan con la media.

```
female_clean["1990"] = pd.to_numeric(female_clean["1990"])
female_clean["2010"] = pd.to_numeric(female_clean["2010"])
female_clean["2018"] = pd.to_numeric(female_clean["2018"])
female_clean["HDI Rank (2018)"] = pd.to_numeric(female_clean["HDI Rank (2018)"])
male_clean["1990"] = pd.to_numeric(male_clean["1990"])
male_clean["2010"] = pd.to_numeric(male_clean["2010"])
male_clean["2018"] = pd.to_numeric(male_clean["2018"])
male_clean["HDI Rank (2018)"] = pd.to_numeric(male_clean["HDI Rank (2018)"])
column_means_female = female_clean.mean()
female_clean = female_clean.fillna(column_means_female)
column_means_male = male_clean.mean()
male_clean = male_clean.fillna(column_means_male)
```

Imagen 4: se pasa de dataframes a float y se llenan datos con la media.

Joining Dataframes

Unimos los dataframes, cambiamos el nombre de las columnas para mayor claridad y eliminamos los duplicados, luego nos aseguramos de que todos los datos numéricos se almacenen como float.

```
[6]  female_clean.rename(columns = {'1990':'1990 F', '2010':'2010 F', '2018':'2018 F'}, inplace = True)
     male_clean.rename(columns = {'1990':'1990 M', '2010':'2010 M', '2018':'2018 M'}, inplace = True)
     dataframe = pd.concat([GDP_clean,male_clean,female_clean], axis=1, join='inner').sort_index()
     dataframe = dataframe.T.drop_duplicates().T
     dataframe.drop(["Country"], axis=1, inplace=True)
     dataframe["1990 F"] = pd.to_numeric(dataframe["1990 F"])
     dataframe["1990 M"] = pd.to_numeric(dataframe["1990 M"])
     dataframe["2010 F"] = pd.to_numeric(dataframe["2010 F"])
     dataframe["2010 M"] = pd.to_numeric(dataframe["2010 M"])
     dataframe["2018 F"] = pd.to_numeric(dataframe["2018 F"])
     dataframe["2018 M"] = pd.to_numeric(dataframe["2018 M"])
     dataframe["Pib"] = pd.to_numeric(dataframe["Pib"])
     dataframe["HDI Rank (2018)"] = pd.to_numeric(dataframe["HDI Rank (2018)"])
```

Imagen 5: Unión de dataframes, se vuelve a pasar de dataframes a float.

Outliers

En este paso se extraen valores atípicos (en términos del PIB) que dificultan el procesamiento de nuestros datos. Además, reacomodamos aleatoriamente las filas para evitar cualquier sesgo durante el procesamiento. También nos deshacemos de las columnas de país y continente dado que no tenían relevancia al momento de agrupar los datos y ademas podría causar inconvenientes.

```
[7]  z = np.abs(stats.zscore(dataframe["Pib"]))
     data = dataframe.copy()
     data["Z_CT"] = z
     data[data["Z_CT"] > 2]
     data = data[data["Z_CT"] <= 2]
     data = shuffle(data)
     dataframe_clean = data.drop(["Pais","Continente","Z_CT"], axis=1)
```

Imagen 6: Se reacomodan aleatoriamente las filas.

Correlation Matrix

A continuación para finalizar el análisis de los datos y limpieza se muestra la matriz de correlación la cual muestra la relación que hay entre los datos que se tienen

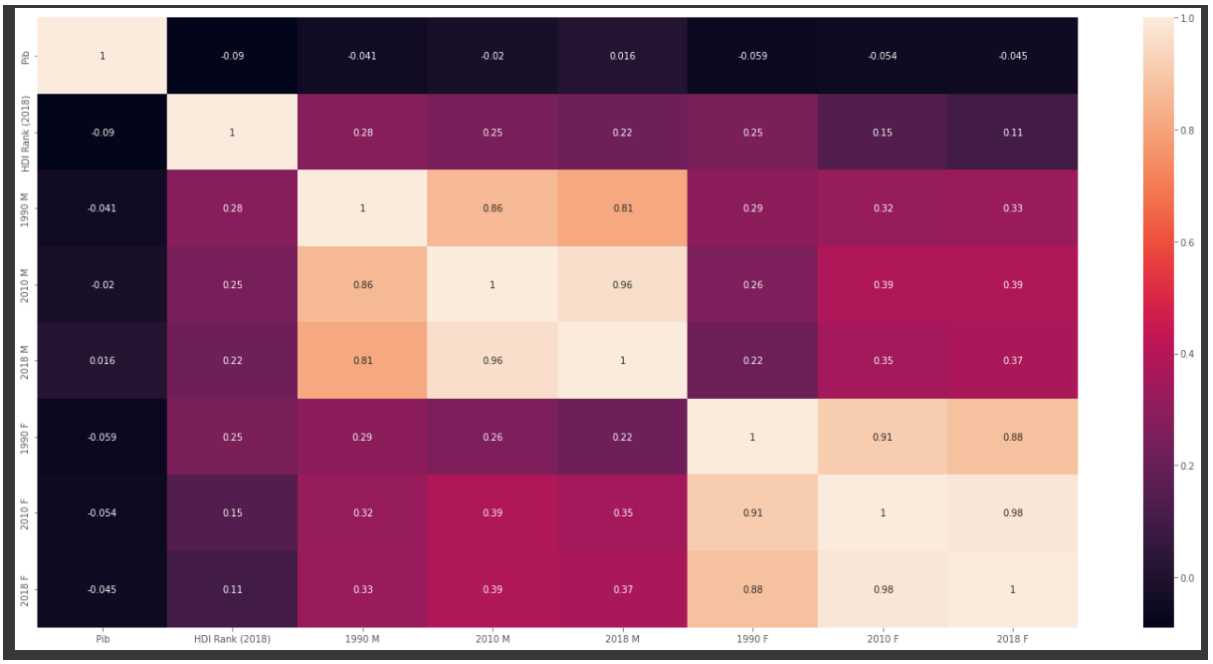

Imagen 7: Matriz de correlación.

PROCESADO DE DATOS

Para el procesado de datos hicimos un análisis de de gráfico de pares para que nos ayudará a comprender un poco cómo se están relacionando las variables en el conjunto de datos

Análisis de gráficos de pares

El objetivo de este análisis es encontrar tres columnas interesantes para crear un análisis adecuado, de modo que podamos separar nuestros datos en grupos encontrando que Aquí, elegimos el PIB, el IDH y el F de 2018.

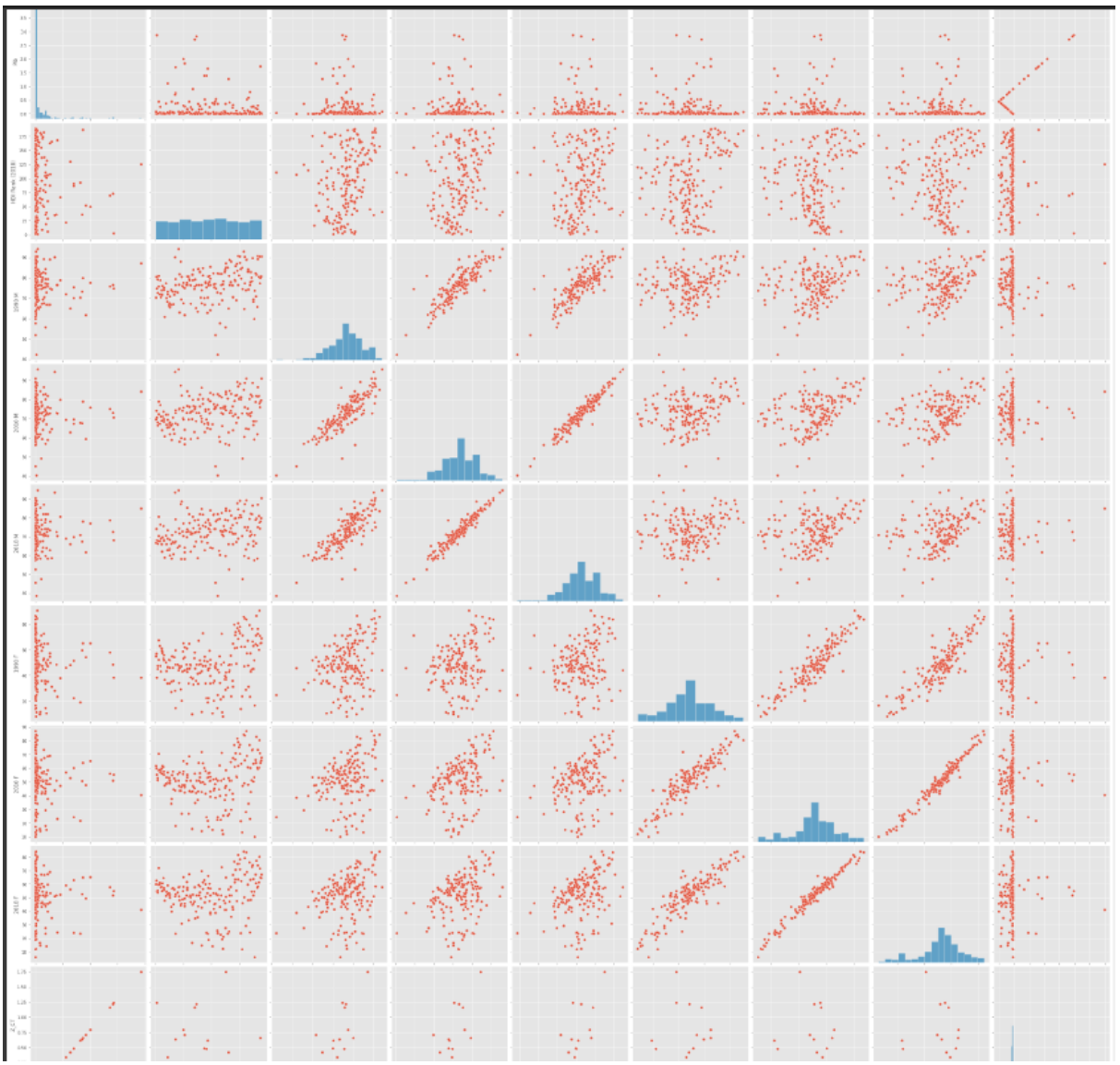

Imagen 8: Diagrama de Pares.

Elbow Curve

Elbow Curve nos permite hacer un agrupamiento que divide un conjunto de observaciones en grupos distintos gracias a valores medios. Pertenece al ámbito de los algoritmos no supervisados

Según las columnas elegidas, Elbow Curve nos ayuda a identificar en cuántos grupos deberíamos agrupar los datos.

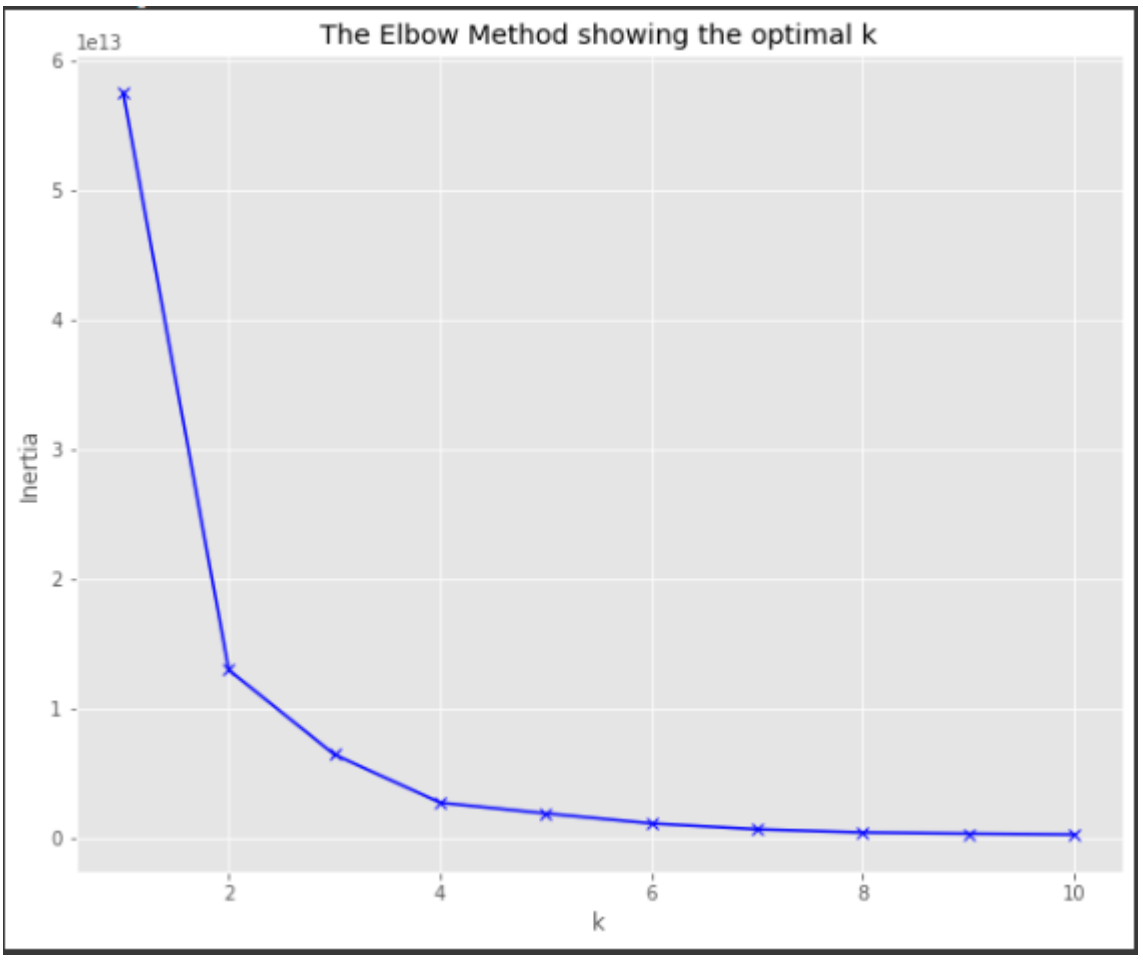

Imagen 7: Curva del codo

La curva aquí claramente se aplana en k = 4, así que ese es el valor que nos da este modelo.

GAP STATISTIC

se implementó Gap Statistic para estimar el número de cluster en nuestro dataset sin embargo no se obtuvo un número parecido al que nos dio Elbow

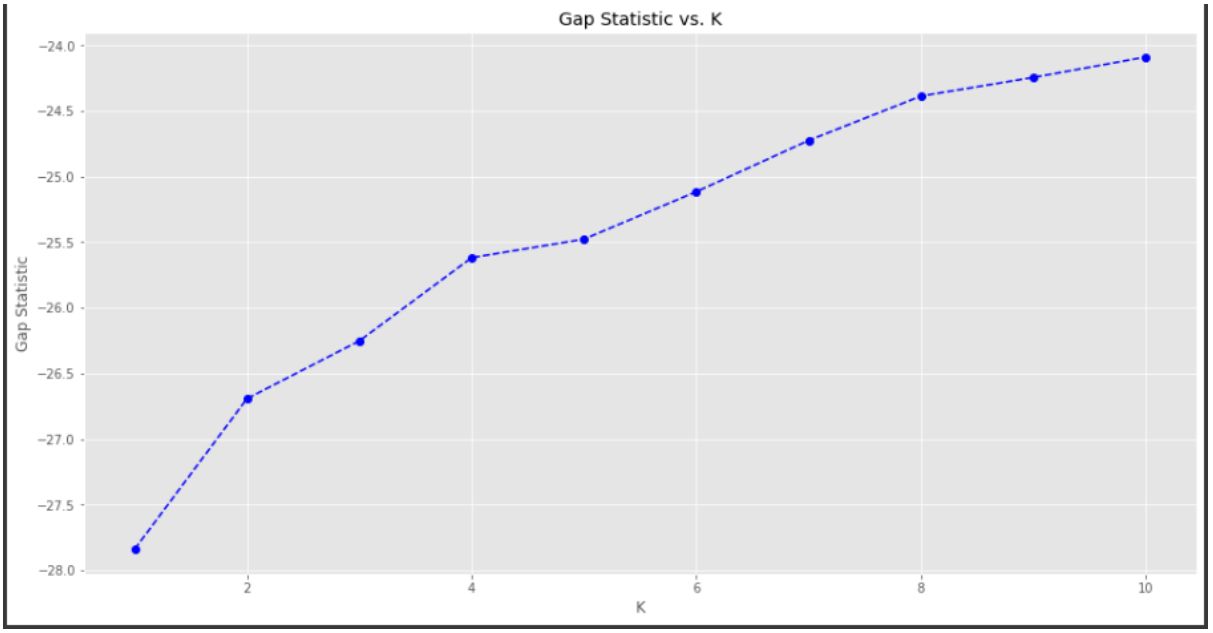

Imagen 10: Curva del codo.

El resultado de 10 es un poco alto, así que buscaremos otro método para confirmar lo que está pasando.

Silhouette Analysis

Este método también se utiliza para determinar los números de conglomerados
El análisis de silueta se puede utilizar para estudiar la distancia de separación entre los grupos resultantes. El gráfico de silueta muestra una medida de qué tan cerca está cada punto en un grupo de puntos en los grupos vecinos y, por lo tanto, proporciona una forma de evaluar parámetros como el número de grupos visualmente.

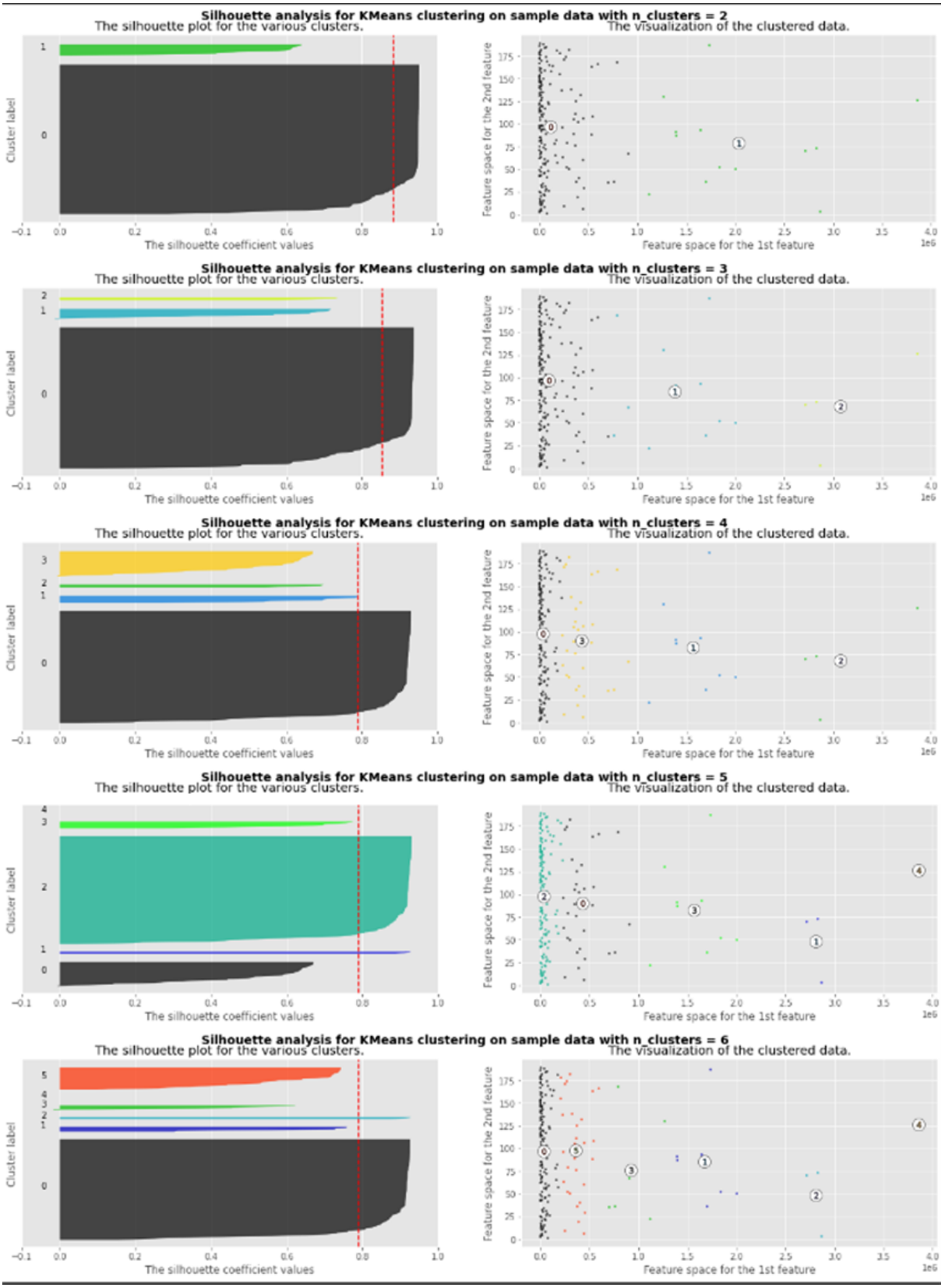

Imagen 11: Análisis de silueta.

Aquí, la puntuación de la silueta se aplana en k = 4, por lo que esto confirma el resultado de la curva del codo.

Dendogram

Un dendrograma es un diagrama que representa un árbol, este realiza un agrupamiento jerárquico de los datos y representa el árbol resultante. Los valores en el eje de profundidad del árbol corresponden a distancias entre grupos.

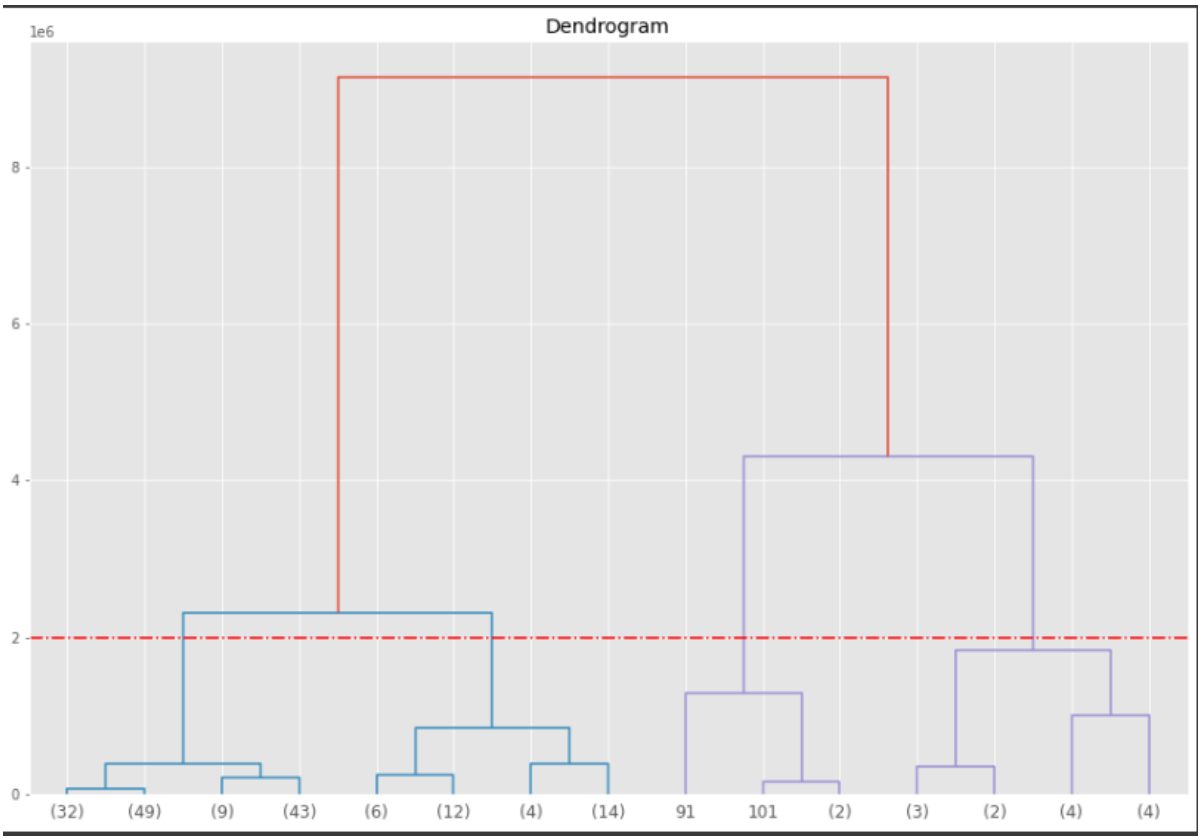

Imagen 12: Dendograma .

Para elegir la altura, tuvimos en cuenta la curva del codo y el análisis de la silueta, dándonos una división en cuatro grupos.

Hierarchical Clustering

Tiene como finalidad crear grupos midiendo las diferencias entre los datos, en nuestro caso los años y los géneros.

	Pib	HDI Rank (2018)	1990 M	2010 M	2018 M	1990 F	2010 F	2018 F
72	0.003700	0.026596	0.744231	0.698556	0.748663	0.720677	0.785714	0.845269
38	0.003165	0.356383	0.801923	0.655235	0.641711	0.299879	0.474026	0.507673
26	0.004129	0.664894	0.828846	0.613718	0.616756	0.413543	0.631169	0.755754
36	0.083803	0.728723	0.546154	0.561372	0.588235	0.685611	0.744156	0.778772
19	0.000643	0.393617	0.407692	0.351986	0.356506	0.401451	0.350649	0.378517

Imagen 13: distribución jerárquica.

Grupos = 4

```
# Create the model with euclidean metric minimizing variability between data
model=AgglomerativeClustering(n_clusters=4, linkage='ward')

#Apply the model
data_fit_4=model.fit(dataframe_clean_copy)
lab_4c=data_fit_4.labels_
dataframe_clean['Labels_4Clusters']=lab_4c
```

Imagen 14.

Caracterización

```
Group 0:
          Pib  HDI Rank (2018)  1990 M  2010 M  2018 M  1990 F  2010 F  \
36   323615.0             138.0    70.6    71.5    71.6    64.8    67.4
24    13469.0             182.0    90.9    81.7    75.1    76.3    59.7
158   54174.0             159.0    90.3    88.9    87.2    84.5    83.9
118  119700.0             158.0    64.7    60.7    59.8    47.1    49.2
92   134628.0             164.0    83.1    75.5    74.9    67.4    60.7

     2018 F  Labels_4Clusters
36     66.9                 0
24     58.5                 0
158    79.4                 0
118    50.6                 0
92     59.8                 0
(44, 9)
```

Imagen 15. Caracterización grupo 0.

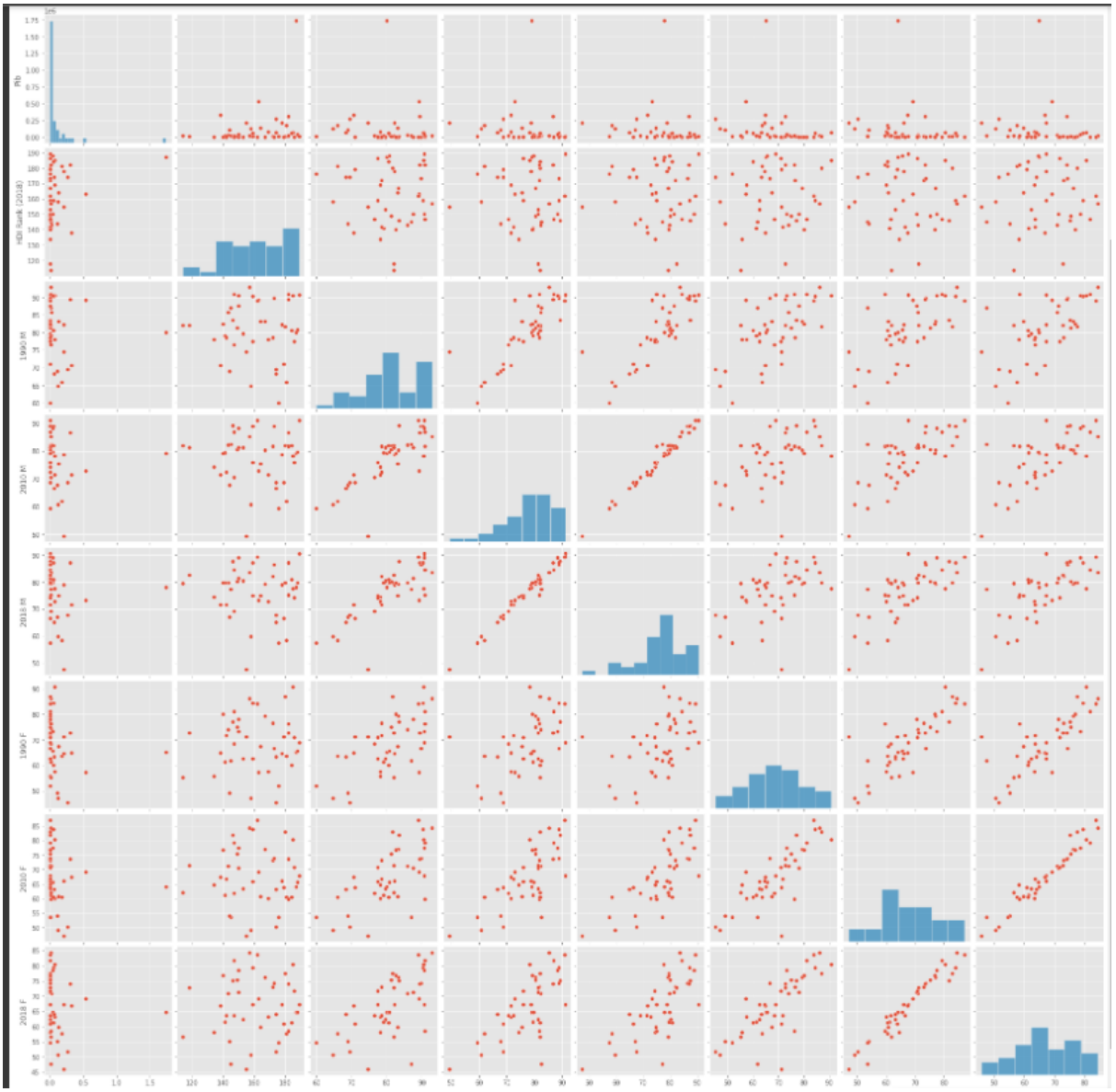

Imagen 16. Diagrama de pair grupo 0.

```
Group 1:
          Pib  HDI Rank (2018)  1990 M  2010 M  2018 M  1990 F  2010 F  \
72    14332.0               6.0    80.9    79.1    80.6    67.7    70.6
42    60752.0              26.0    70.9    68.0    68.4    52.0    49.2
14     5209.0              17.0    60.8    60.8    58.9    36.6    47.5
79   453996.0              29.0    66.9    59.0    58.4    35.3    37.8
171    8116.0              15.0    75.3    69.9    68.2    56.2    57.5

     2018 F  Labels_4Clusters
72     72.1                 1
42     52.4                 1
14     47.9                 1
79     40.0                 1
171    56.1                 1
(53, 9)
```

Imagen 17. Caracterización grupo 1.

```
Group 2:
          Pib  HDI Rank (2018)  1990 M  2010 M  2018 M  1990 F  2010 F  \
137   38145.0       111.000000    42.2    40.4    38.6    24.9    24.3
140  595858.0       166.000000    74.4    64.7    58.6    31.4    34.3
2    171091.0        82.000000    76.6    70.1    67.4    11.4    14.4
147    2122.0        94.772487    76.2    74.9    74.3    22.3    17.6
157  372062.0       125.000000    55.8    57.0    59.7    29.0    29.5

     2018 F  Labels_4Clusters
137    23.7                 2
140    35.2                 2
2      14.9                 2
147    19.1                 2
157    27.8                 2
(32, 9)
```

Imagen 18: Caracterización grupo 2.

```
Group 3:
          Pib  HDI Rank (2018)     1990 M     2010 M     2018 M     1990 F  \
189   57921.0             55.0  76.451667  73.072778  72.572222  48.492222
38    12267.0             68.0  83.900000  76.700000  74.600000  32.900000
159   54174.0             77.0  84.600000  80.100000  76.200000  67.200000
88    95503.0            122.0  74.500000  76.600000  75.800000  50.200000
46   350104.0             89.0  77.400000  74.000000  77.600000  33.800000

        2010 F     2018 F  Labels_4Clusters
189  50.959444  51.811667                 3
38   46.600000  45.700000                 3
159  63.900000  59.500000                 3
88   52.200000  48.000000                 3
46   41.600000  50.900000                 3
(57, 9)
```

Imagen 19: Caracterización grupo 3.

Resultados

Grupo 0

	Pib	HDI Rank (2018)	1990 M	2010 M	2018 M	1990 F	2010 F	2018 F
count	44	44	44	44	44	44	44	44
mean	106476	161	81	78	76	69	68	67
std	275240	19	8	9	9	11	10	10
min	194	114	60	49	48	45	47	46
25%	2853	147	77	72	72	62	61	61
50%	14219	160	82	80	79	69	66	67
75%	84391	178	88	82	82	76	74	74
max	1736425	189	93	91	90	91	87	84

Imagen 19: Resultados grupo 0.

Grupo 1

	Pib	HDI Rank (2018)	1990 M	2010 M	2018 M	1990 F	2010 F	2018 F
count	53	53	53	53	53	53	53	53
mean	501611	36	72	68	68	50	53	54
std	900825	30	5	6	6	9	7	7
min	268	1	61	57	58	27	34	40
25%	14989	14	68	64	64	46	49	50
50%	51475	28	73	68	67	51	53	55
75%	445445	50	76	71	70	55	58	58
max	3861123	130	87	84	85	68	71	72

Imagen 20: Resultados grupo 1.

Grupo 2

	Pib	HDI Rank (2018)	1990 M	2010 M	2018 M	1990 F	2010 F	2018 F
count	32	32	32	32	32	32	32	32
mean	194340	113	72	68	68	25	26	27
std	398800	34	10	10	10	11	10	11
min	118	36	42	40	39	8	10	6
25%	3723	92	67	65	63	16	18	21
50%	16732	114	74	69	70	24	25	24
75%	165374	132	79	75	73	32	35	36
max	1646739	177	88	83	89	46	45	49

Imagen 21:Resultados grupo 3.

Grupo 3

	Pib	HDI Rank (2018)	1990 M	2010 M	2018 M	1990 F	2010 F	2018 F
count	57	57	57	57	57	57	57	57
mean	133314	91	80	78	77	44	50	52
std	192636	33	5	5	6	9	8	7
min	47	35	70	67	66	20	27	34
25%	12296	66	76	74	73	38	45	48
50%	52091	89	79	76	76	45	51	52
75%	181665	108	83	80	81	48	54	57
max	907050	171	94	96	95	73	72	70

Imagen 22: Resultados grupo 3.

RETOS

- nuestro mayor reto fue tener que cambiar el dataset ya que encontramos que no había correlación con el precio y con la marca y además la marca era escogida por el usuario lo que no nos permite implementar el modelo
- también el primer Dataset era demasiado grande lo que se dificulta bastante a la hora procesar
- sobre el modelo que implementamos uno de los mayores desafíos fue encontrar a forma de agrupar los datos para que brindara información relevante de cada grupo de países

CONCLUSIONES

- Dado que los 4 grupos de países tienen en común que con el paso de los años las mujeres han incrementado bastante su papel en el área laboral lo que es un indicativo de que se están desarrollando nuevas políticas de inclusión laboral además de una disminución de la discriminacion laboral hacia las mujeres

- podemos concluir que a mayor porcentaje de fuerza laboral el hdi rank incrementa

- Podemos que el pib no siempre se observe directamente con el porcentaje de fuerza laboral, ya que en el grupo dos que es el de mayor porcentaje no obstante tiene el pib más alto, por lo que concluimos que falta mucho más factores para calcularlo.

- Los países con grandes porcentajes de fuerza laboral, ayudan bastante a que un grupo destaque en su media, que nos fijamos que normalmente el máximo de cada grupo influye considerablemente.

  A pesar de que la tendencia sea que el porcentaje de fuerza laboral femenina aumente, notamos que para todos los grupos el porcentaje masculino sigue siendo mayor.