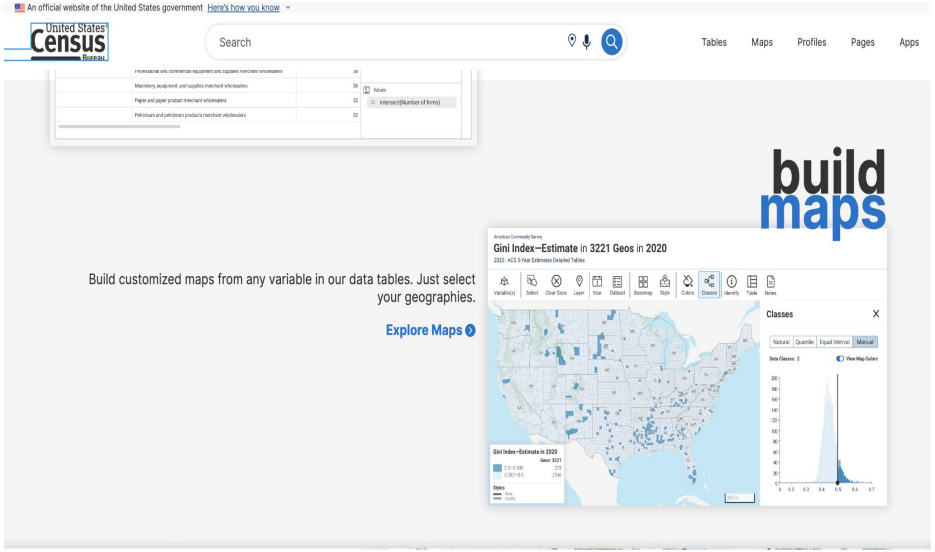
Census Income

Alejandro Pallais - Rudik Rompich

Datos

- Este conjunto de datos fue extraído de la base de datos del Census Bureau de EEUU.
- Contiene 32561 registros. Estamos usando una muestra de la base de datos completa. [Kaggle]
- Incluye una mezcla de datos continuos y discretos.

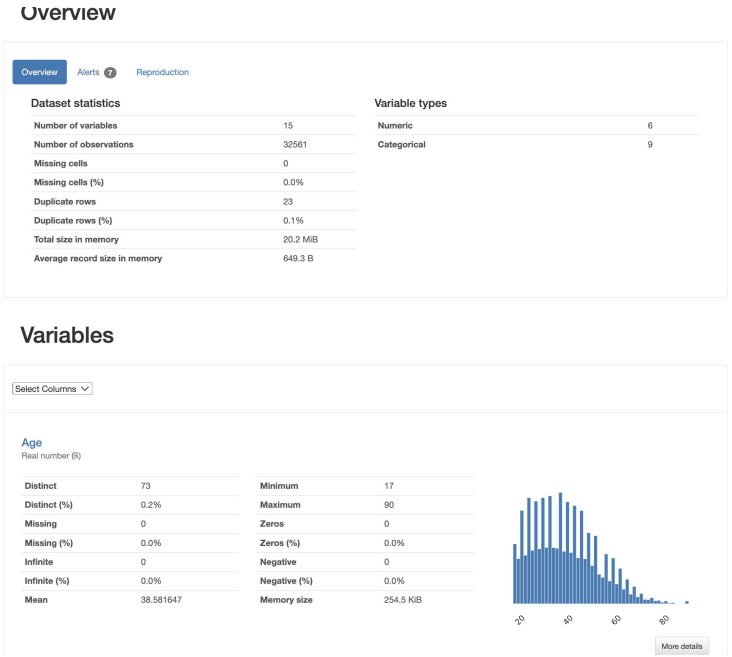


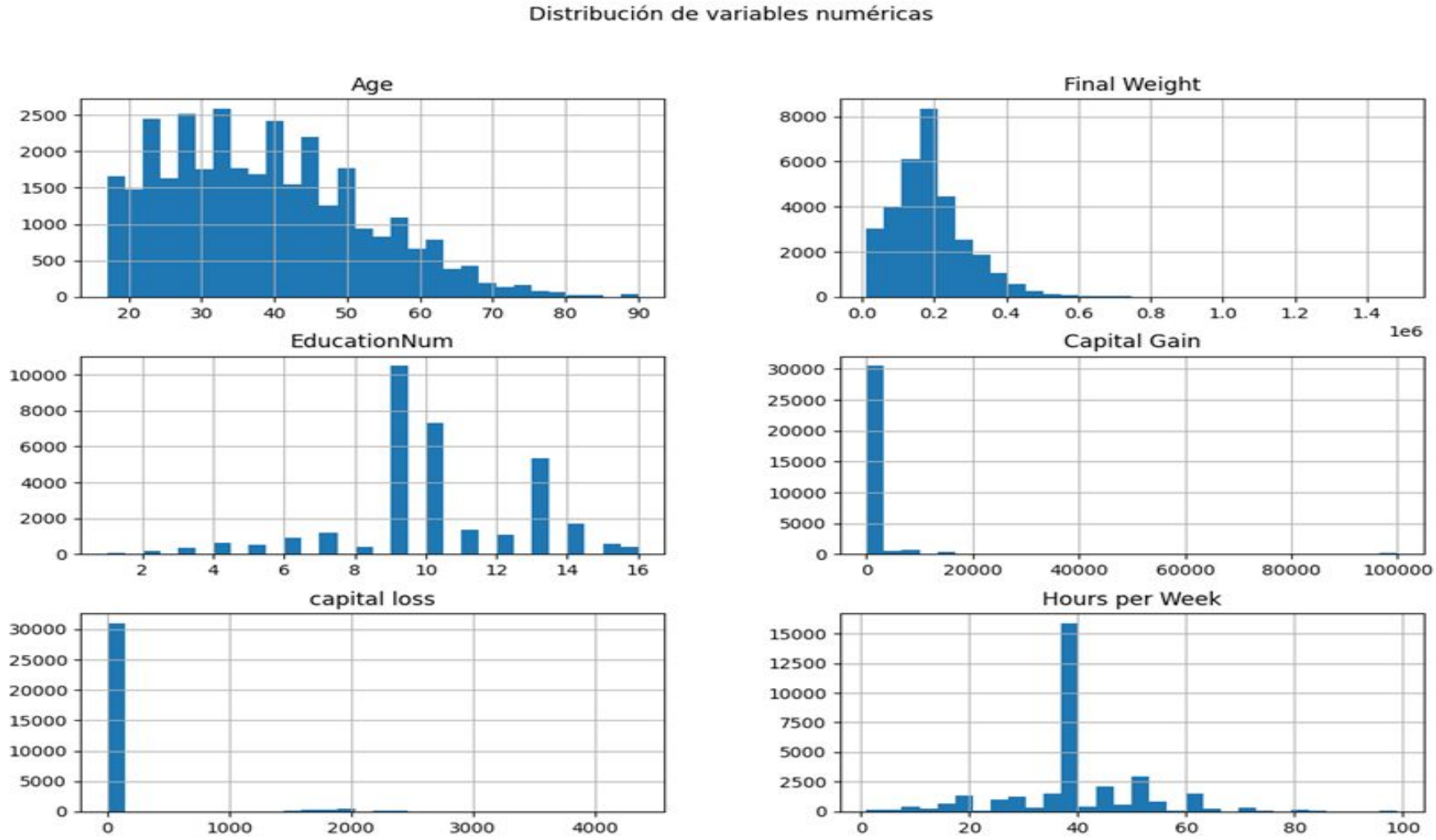
Datos

- **age**: Edad de la persona.
- **workclass**: Tipo de trabajo o empleo, categorizado en opciones como sector privado, autónomos, gobierno federal, local y estatal, entre otros.
- **Final Weight**: Ponderación del individuo en el archivo CPS, ajustado según diversas estimaciones demográficas.
- **education**: Nivel educativo más alto alcanzado.
- **education-num**: Número de años de educación completados.
- **marital-status**: Estado civil de la persona.
- **occupation**: Tipo de ocupación o trabajo realizado.
- **relationship**: Estado de la relación (por ejemplo, cónyuge, hijo, etc.).
- **race**: Raza de la persona.
- **sex**: Género de la persona.
- **capital-gain**: Ganancia de capital obtenida.
- **capital-loss**: Pérdida de capital incurrida.
- **hours-per-week**: Número de horas trabajadas por semana.
- **native-country**: País de origen o país natal.
- **income**: Nivel de ingresos, indicando si son mayores a $50,000 o menores o iguales a $50,000.

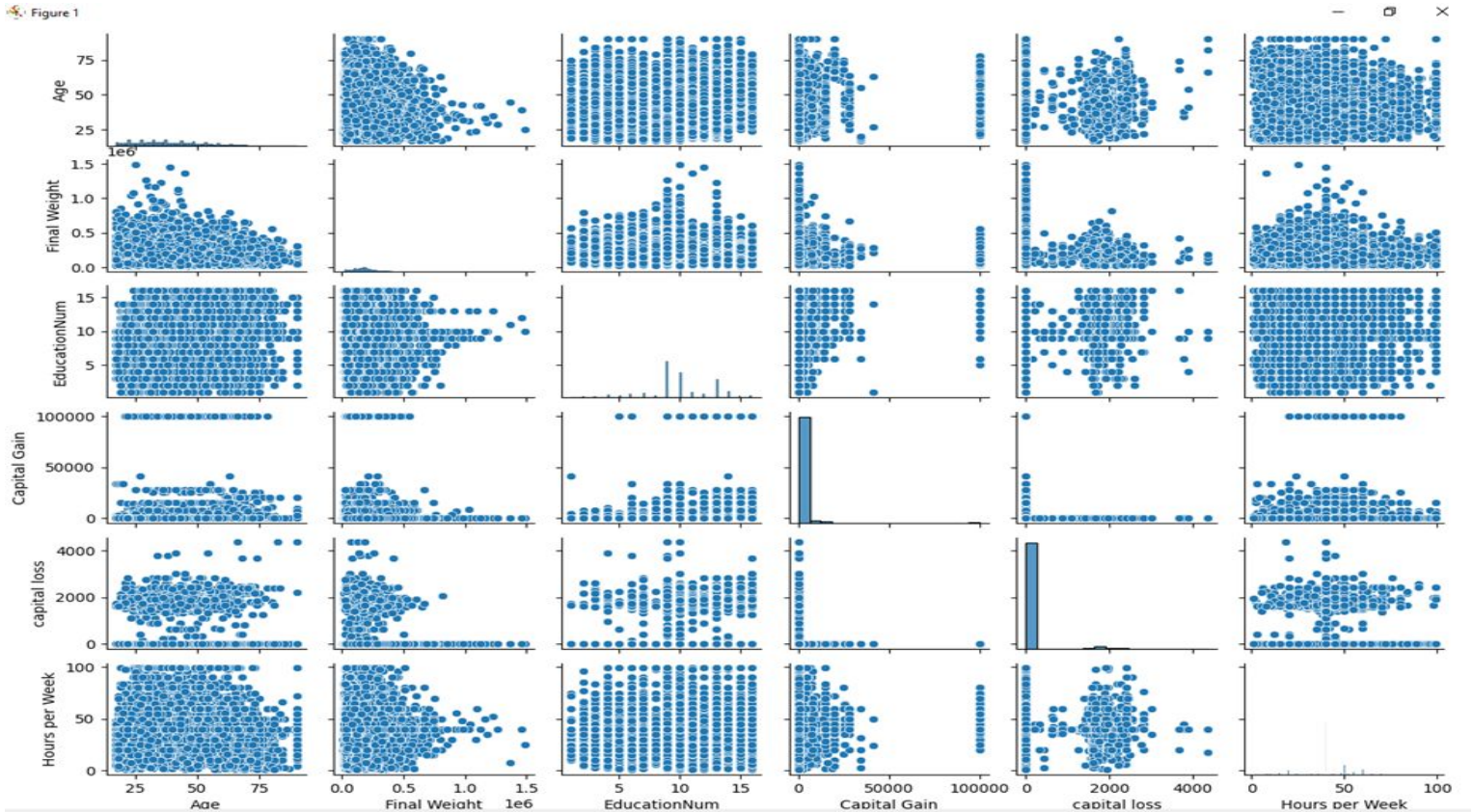
Datos

- **Pandas profiling**



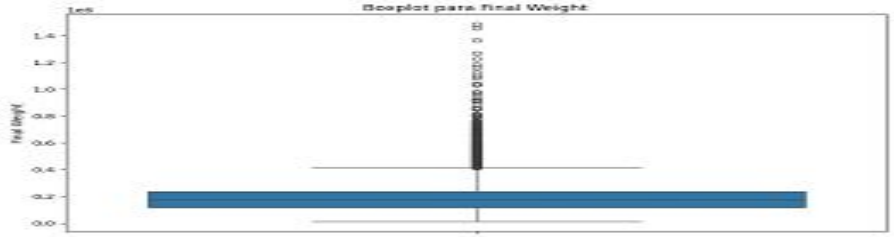

Distribución de variables numéricas
Age
Final Weight
1e6
EducationNum
Capital Gain
capital loss
Hours per Week

Gráficos de dispersión

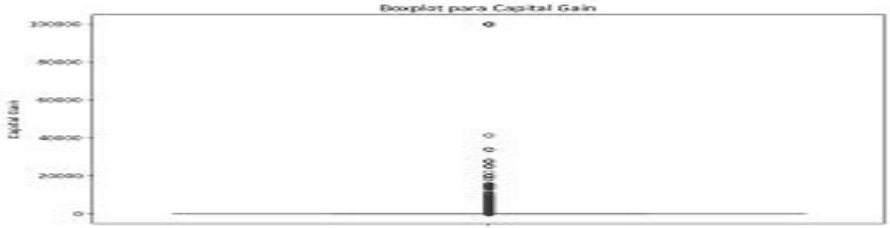


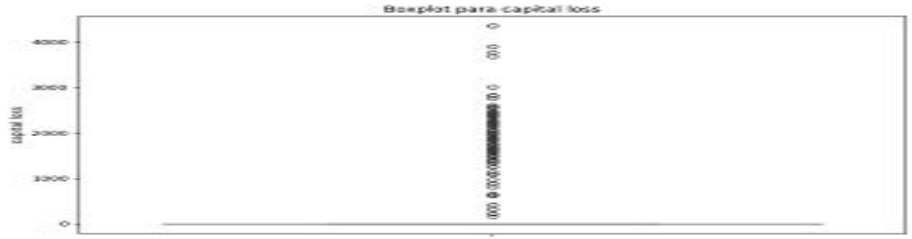
boxplots

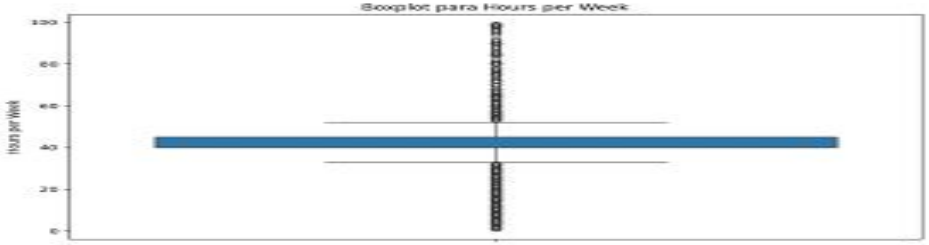

Boxplot para Age
Age


Boxplot para final Weight
Fnal Weight

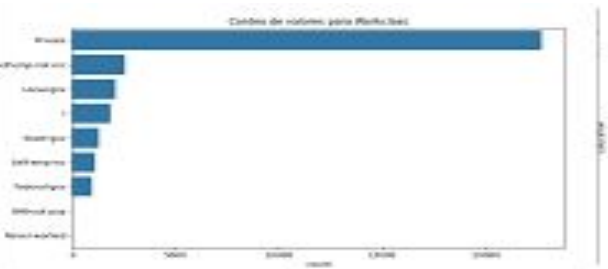

Boxplot para EducationNum

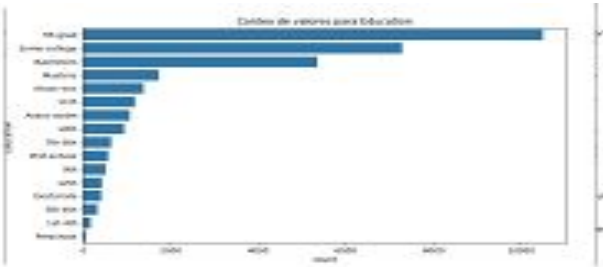

Boxplot para Capital Gain
Capital Gain

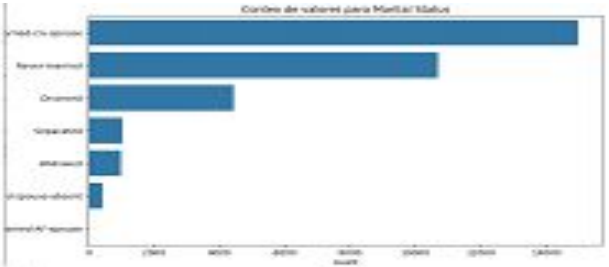

Boxplot para capital loss
capital loss

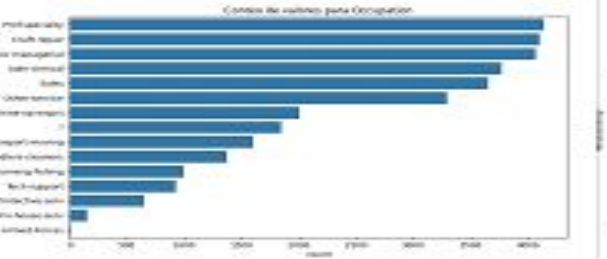

Boxplot para Hours per Week
Hours per Week

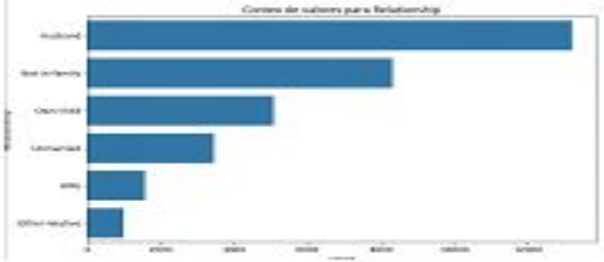
Conteo de valores para variables categóricas

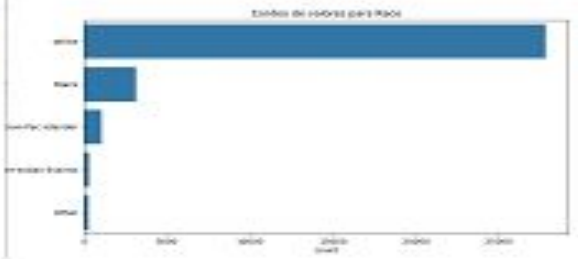


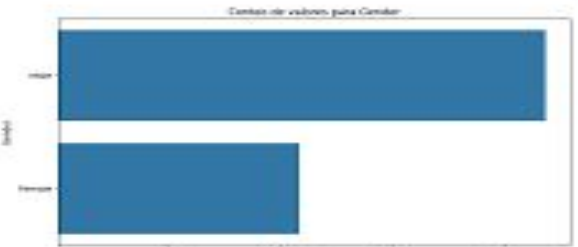


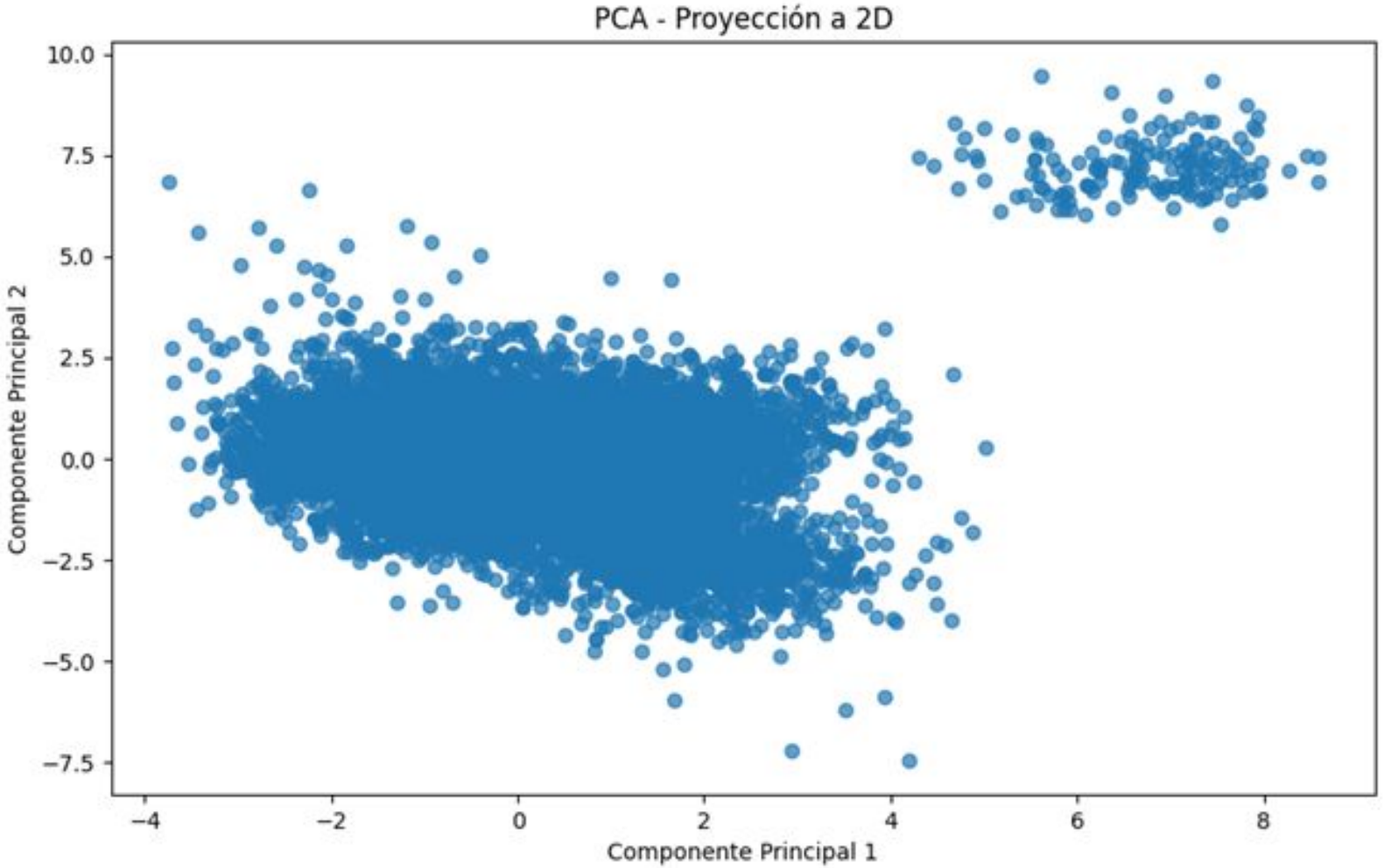


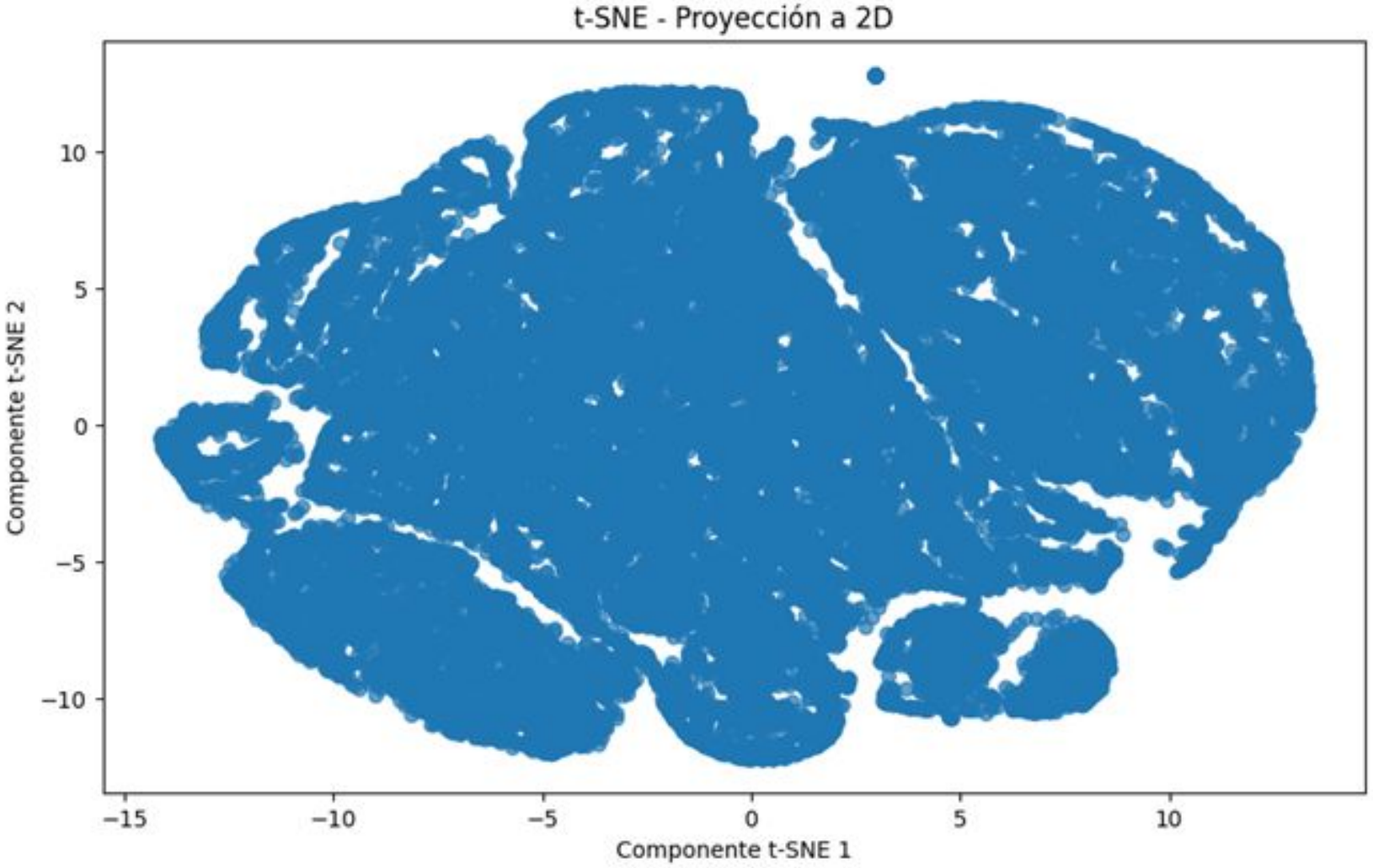


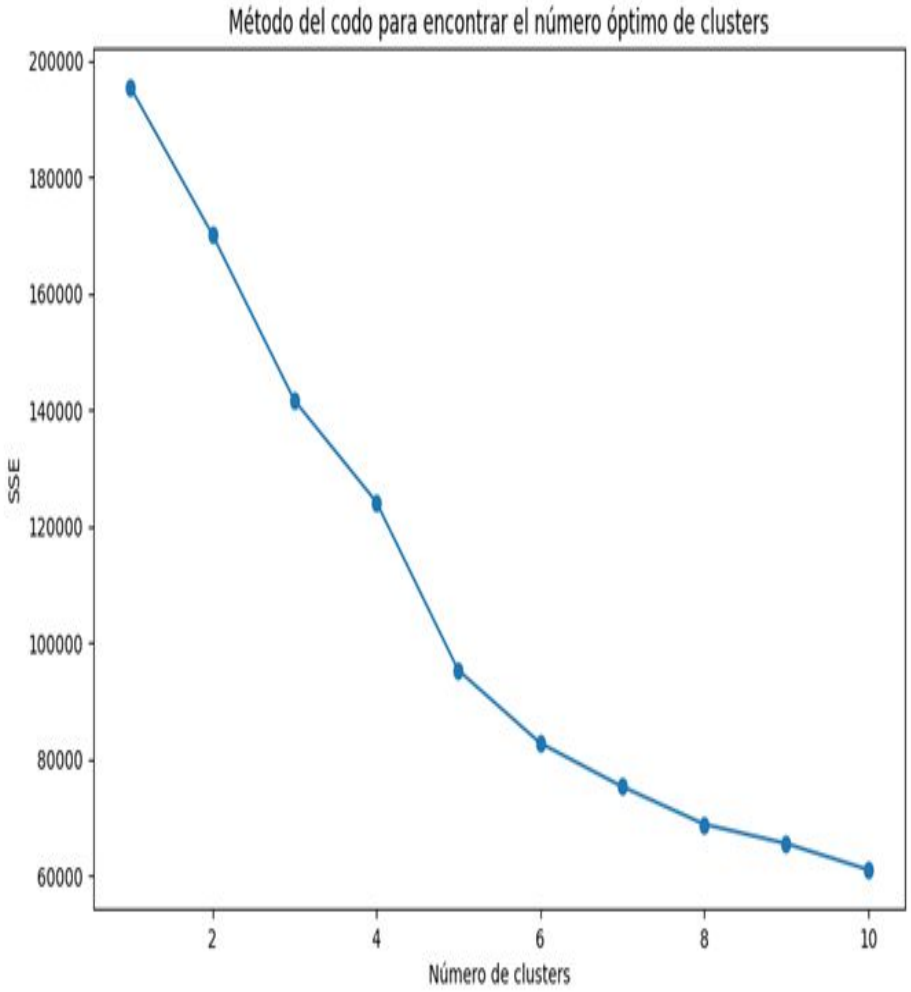


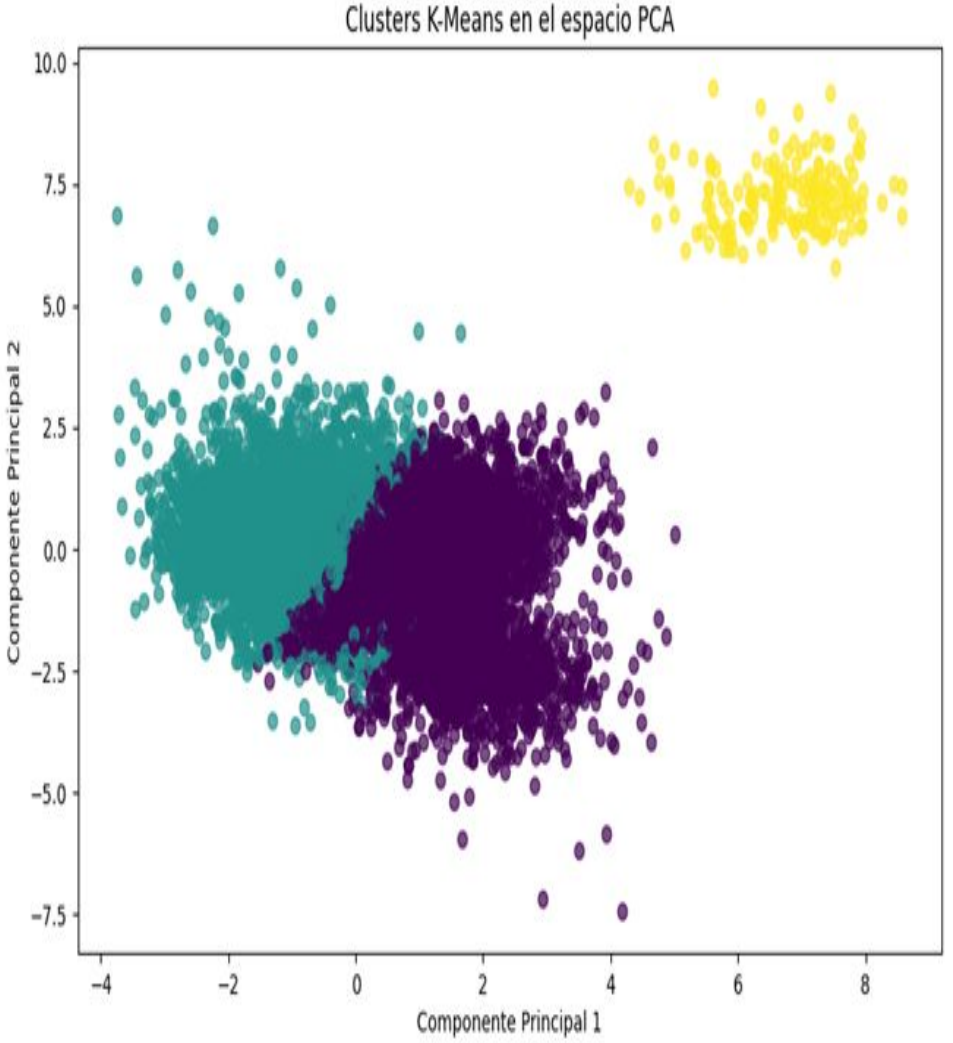


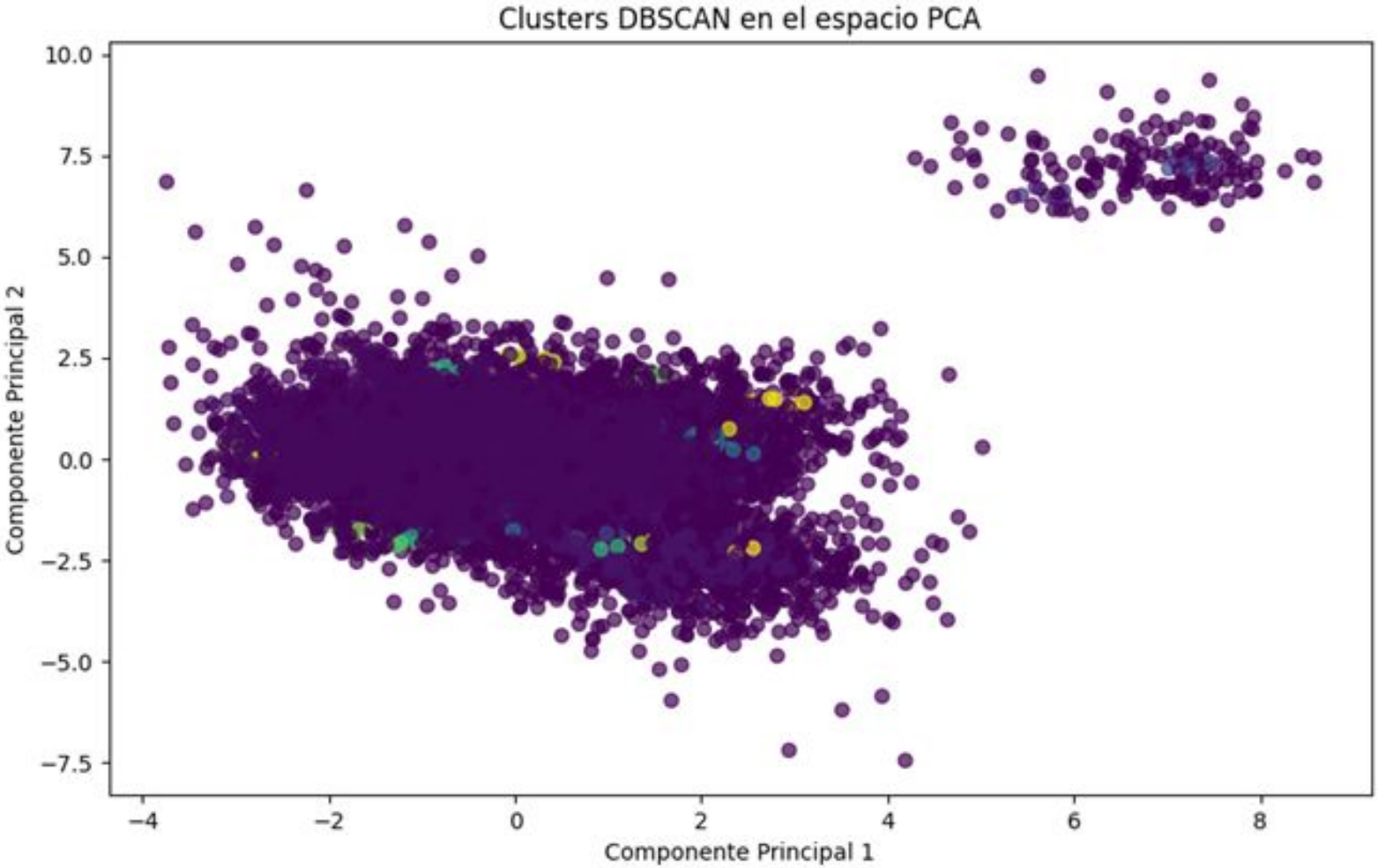


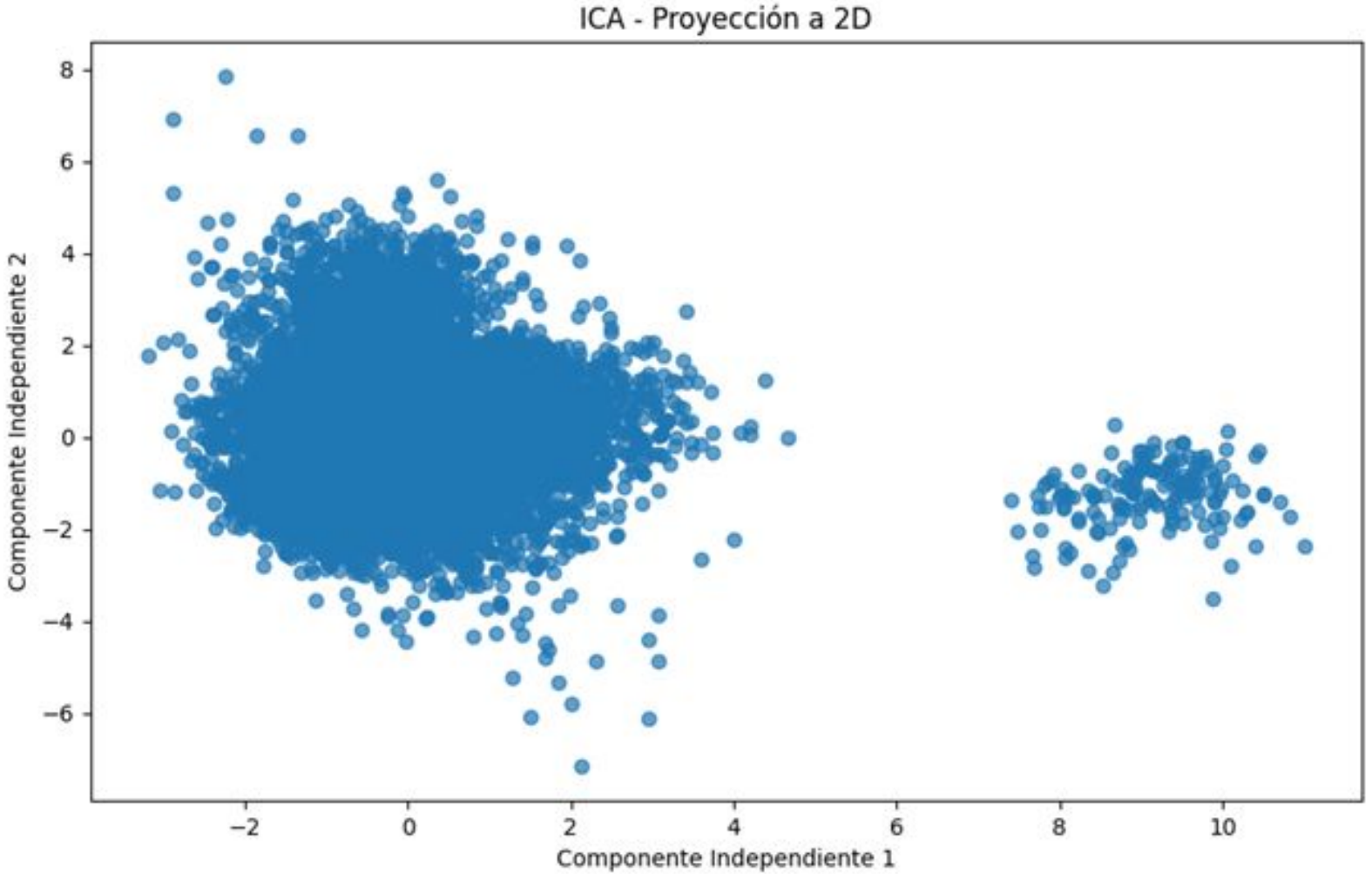

PCA - Proyección a 2D
Componente Principal 2
Componente Principal 1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−4
−2
0
2
4
6
8

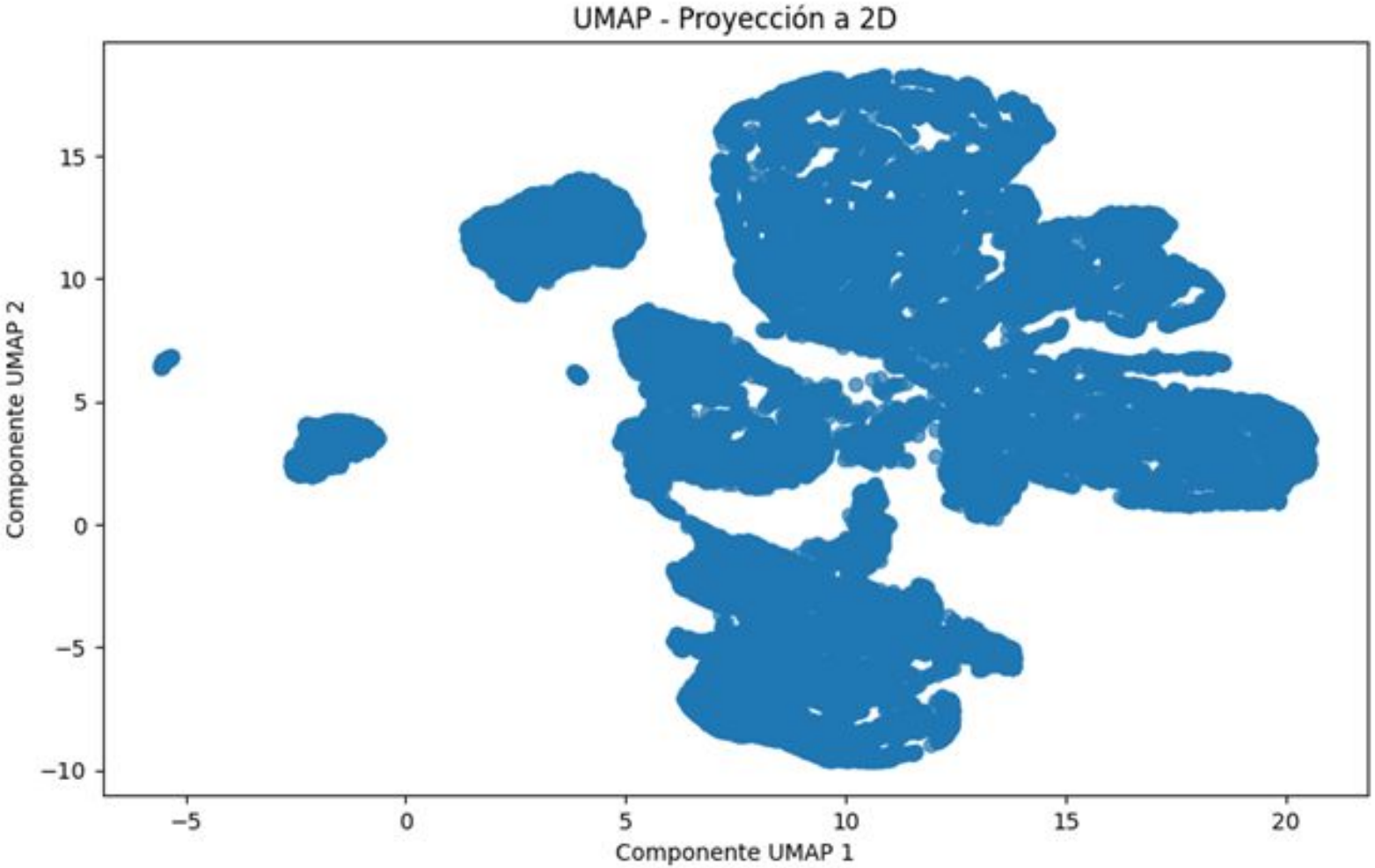

t-SNE - Proyección a 2D
Componente t-SNE 2
Componente t-SNE 1
10
5
0
−5
−10
−15
−10
−5
0
5
10


Método del codo para encontrar el número óptimo de clusters
SSE
Número de clusters
200000
180000
160000
140000
120000
100000
80000
60000
2
4
6
8
10


Clusters K-Means en el espacio PCA
Componente Principal 2
Componente Principal 1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−4
−2
0
2
4
6
8

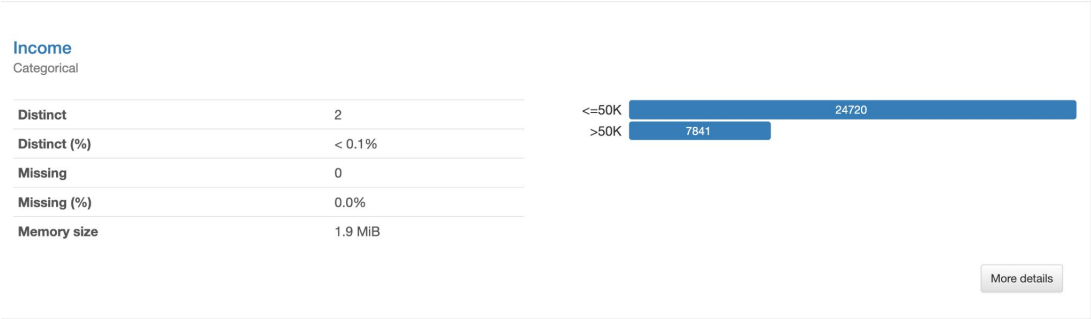

Clusters DBSCAN en el espacio PCA
Componente Principal 2
Componente Principal 1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−4
−2
0
2
4
6
8


ICA - Proyección a 2D
Componente Independiente 2
Componente Independiente 1
8
6
4
2
0
−2
−4
−6
−2
0
2
4
6
8
10

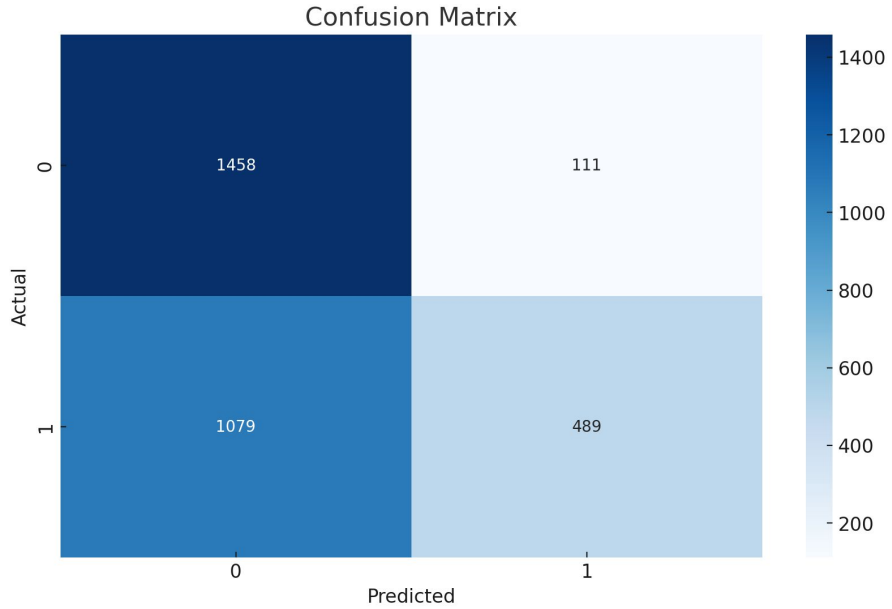

UMAP - Proyección a 2D
Componente UMAP 2
Componente UMAP 1
15
10
5
0
−5
−10
−5
0
5
10
15
20

Regresión Logística

- OLS
- No multicolinealidad
- Problema con balanceo
- 80%/20%



Regresión Logística



SVM (Support Vector Machine)

	precision	recall	f1-score	support
<=50K	0.88	0.94	0.91	7455
>50K	0.74	0.58	0.65	2314
accuracy			0.85	9769
macro avg	0.81	0.76	0.78	9769
weighted avg	0.85	0.85	0.85	9769

- 70% / 30%