Proyecto_final

July 1, 2025

Universidad Autónoma Metropolitana - Unidad Iztapalapa (UAM-I)

Maestría en Matemáticas Aplicadas e Industriales (MCMAI)

Taller de Modelado Matemático II - Parte I

Trimestre 25-P

Profesor:

Dr. Alejandro Román Vásquez

Alumnos:

Alan Badillo Salas

Brandon Eduardo Antonio Gómez

1 Proyecto Final

2 Fase 1 - Adquisición de los datos

Cargamos el dataset "Adult" notando que faltan los encabezados (cabeceras) y hay explicar quiénes son las columnas (sacadas de `adult.names`)

```
   age         workclass  fnlwgt  education  education-num  \
0   39         State-gov   77516  Bachelors             13
1   50  Self-emp-not-inc   83311  Bachelors             13
2   38           Private  215646    HS-grad              9
3   53           Private  234721       11th              7
4   28           Private  338409  Bachelors             13

       marital-status         occupation   relationship   race     sex  \
0       Never-married       Adm-clerical  Not-in-family  White    Male
1  Married-civ-spouse    Exec-managerial        Husband  White    Male
2            Divorced  Handlers-cleaners  Not-in-family  White    Male
3  Married-civ-spouse  Handlers-cleaners        Husband  Black    Male
4  Married-civ-spouse     Prof-specialty           Wife  Black  Female

   capital-gain  capital-loss  hours-per-week native-country  income
0          2174             0              40  United-States   <=50K
```

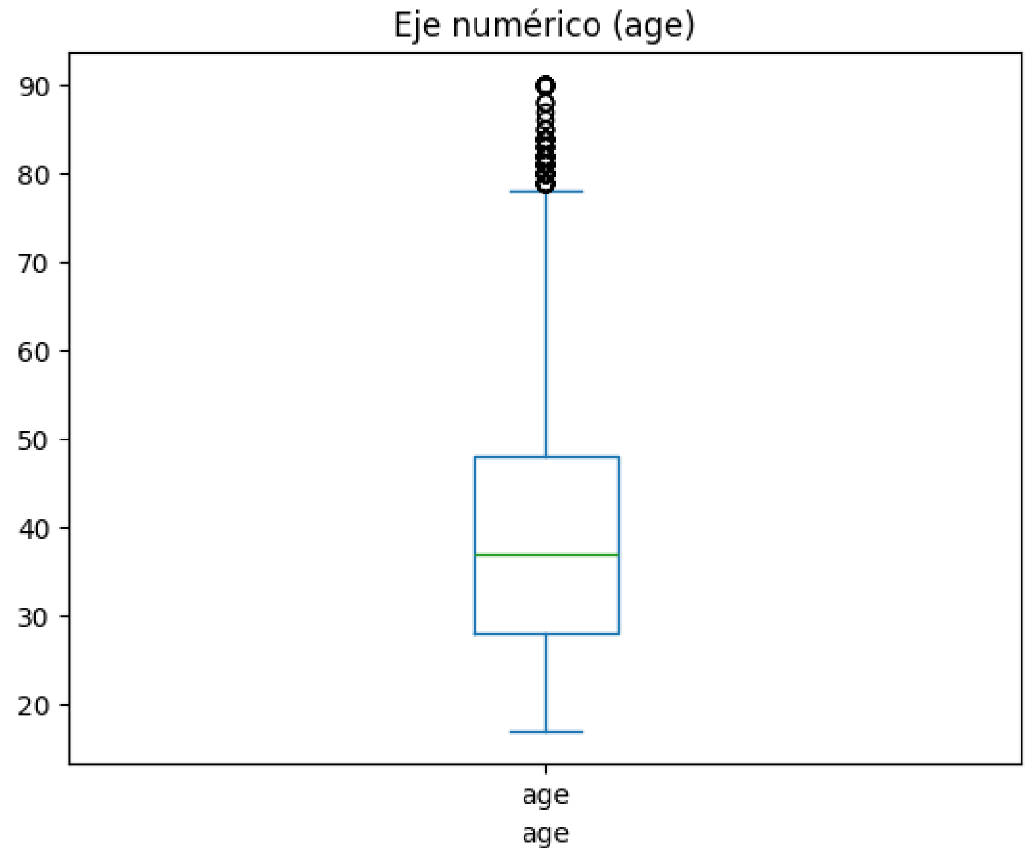
```
1                0                0                13   United-States   <=50K
2                0                0                40   United-States   <=50K
3                0                0                40   United-States   <=50K
4                0                0                40            Cuba   <=50K
```

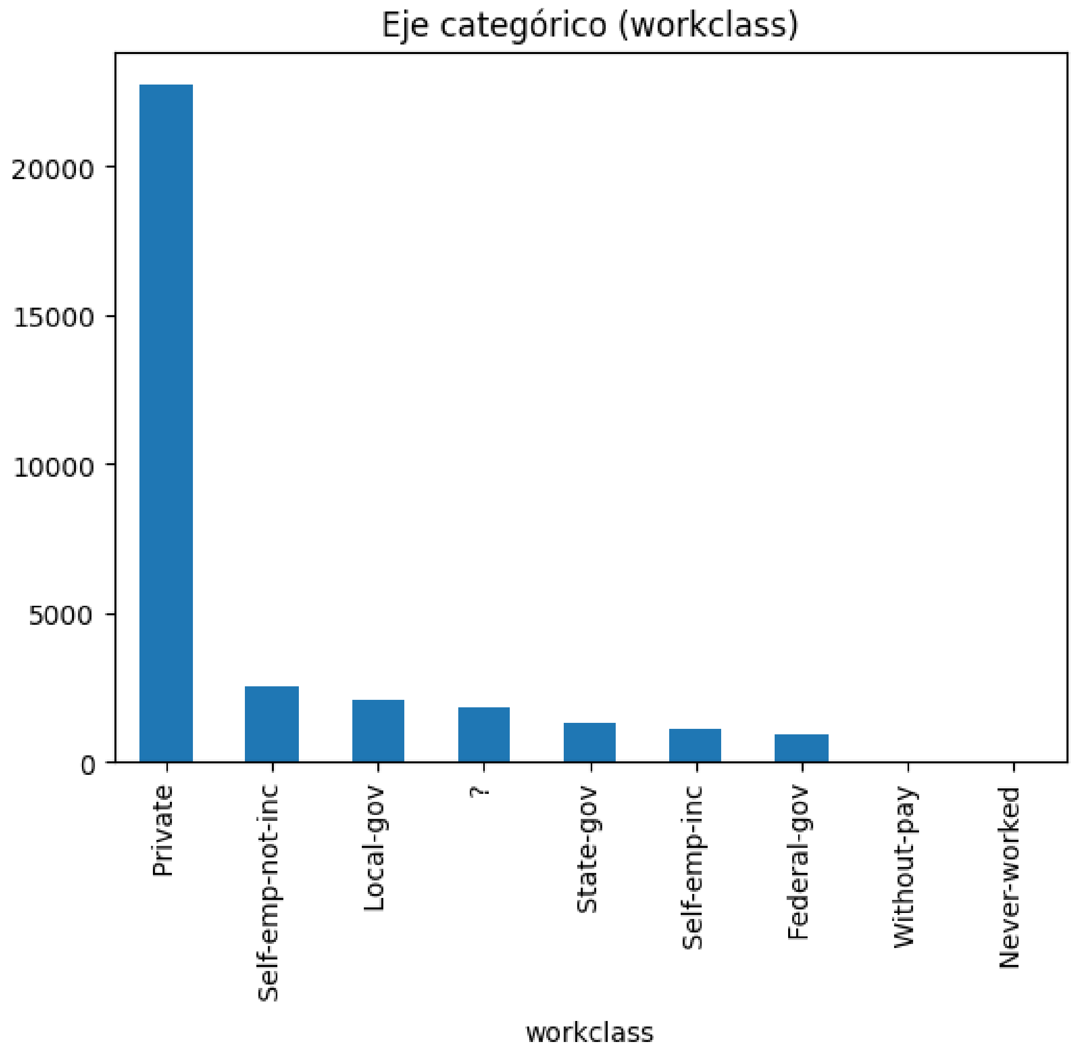
Analizamos la información general observando 32,561 registros en 15 columnas no nulas.

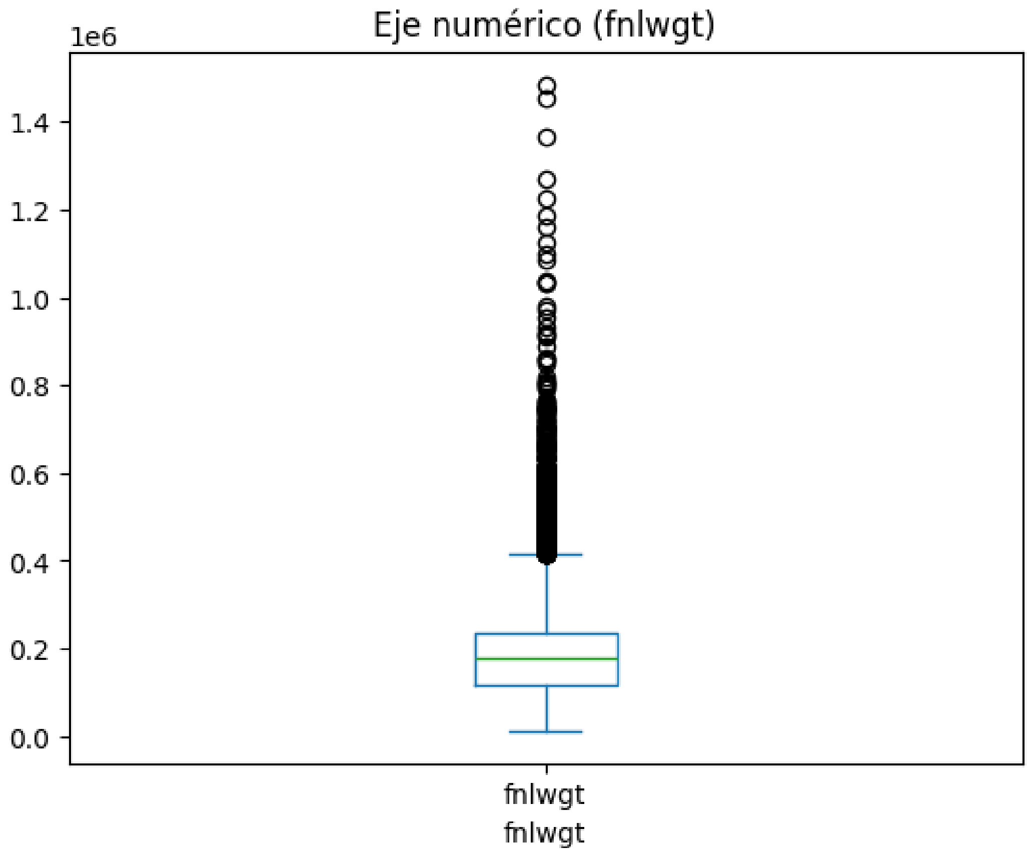
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  native-country  32561 non-null  object
 14  income          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

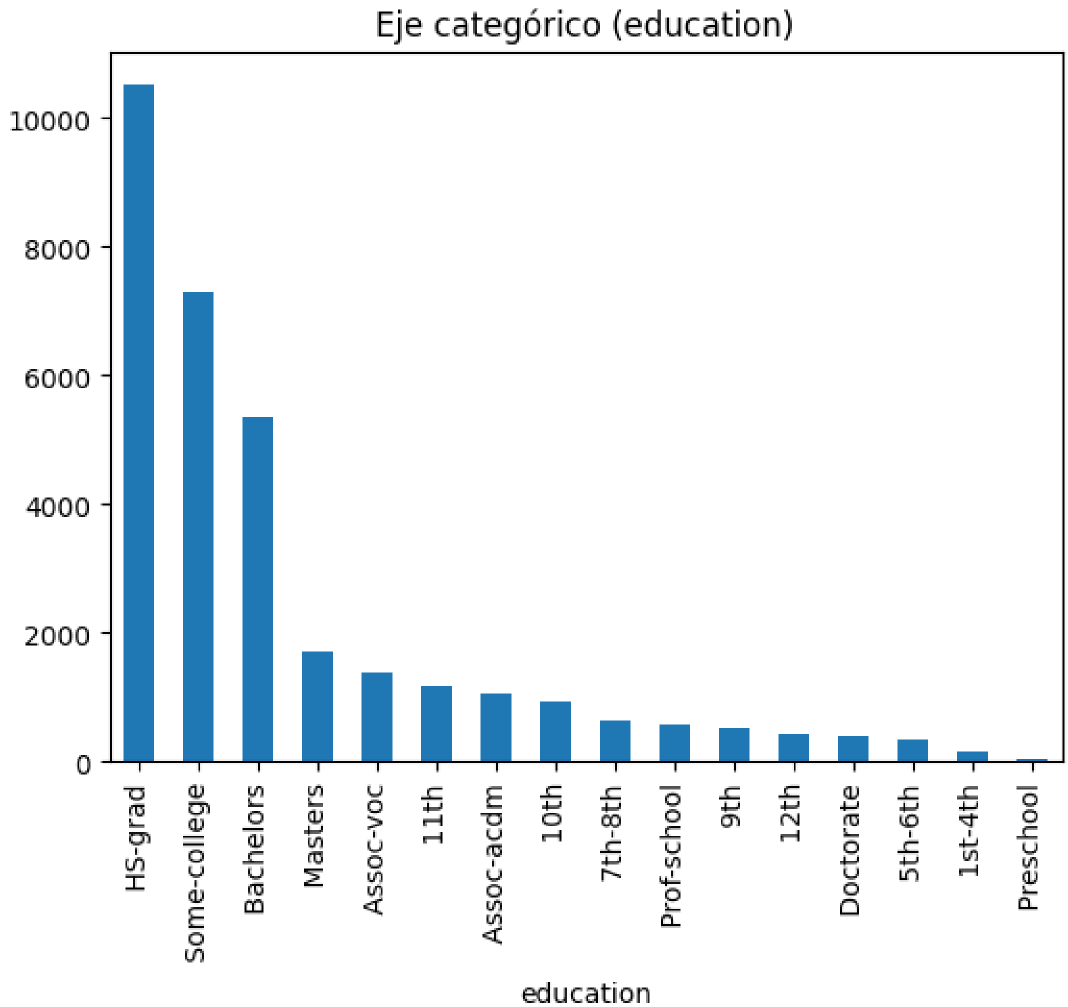
3 Fase 2 - Ingeniería de variables

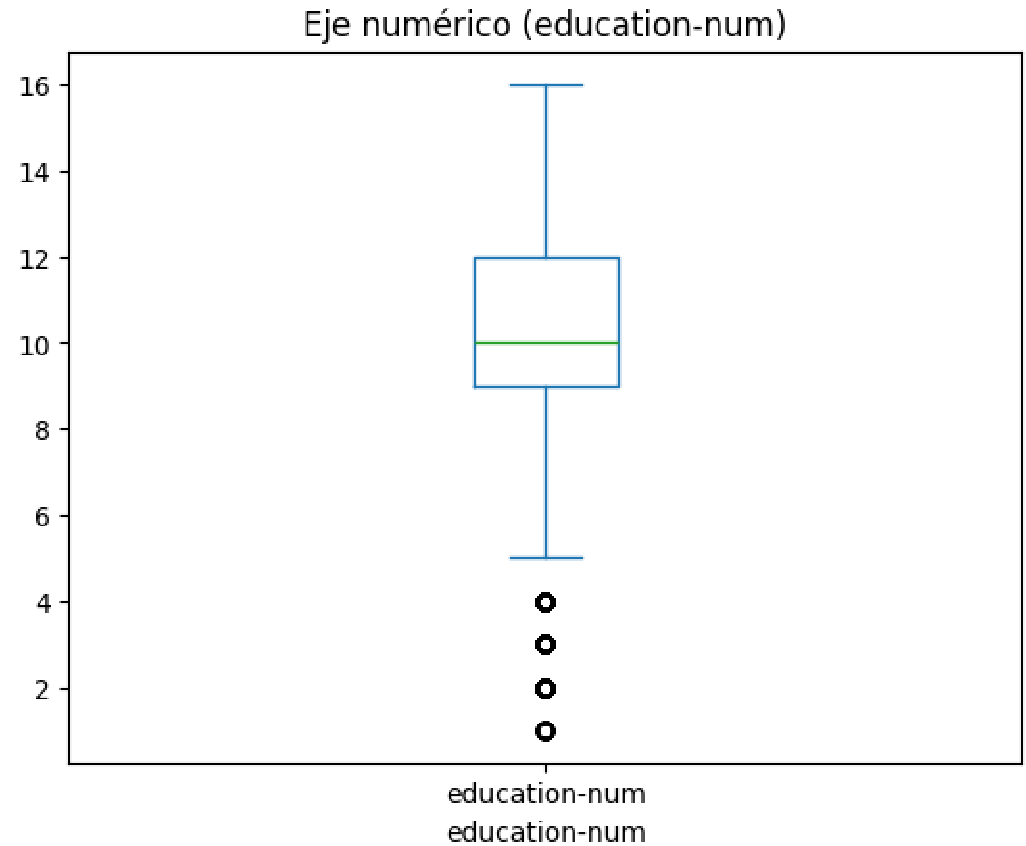
Primero extraemos los ejes de datos observando que hay 6 numéricos y 9 categóricos que analizaremos para extraer las posibles variables de análisis.

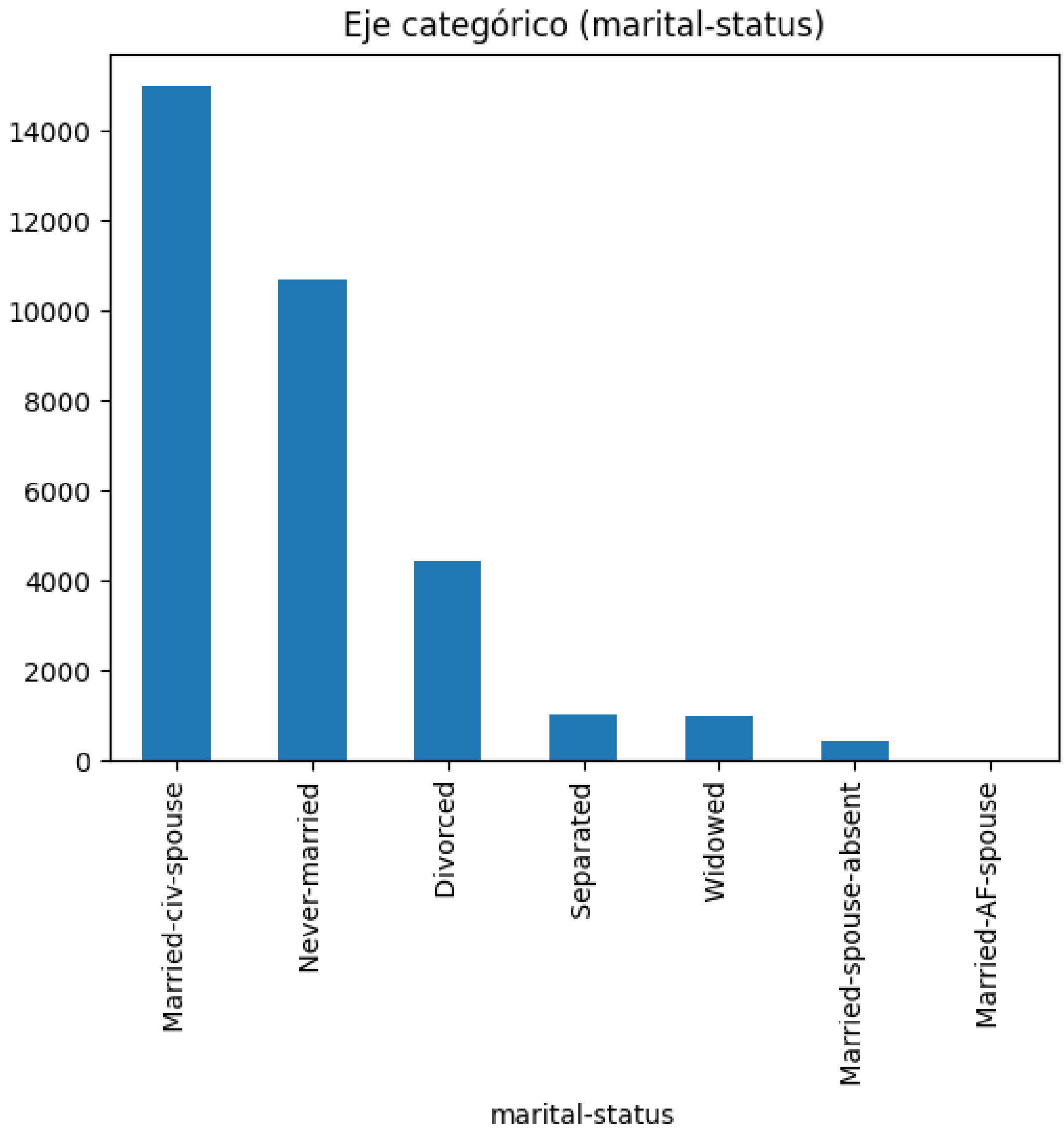

Eje numérico (age)
90
80
70
60
50
40
30
20
age
age

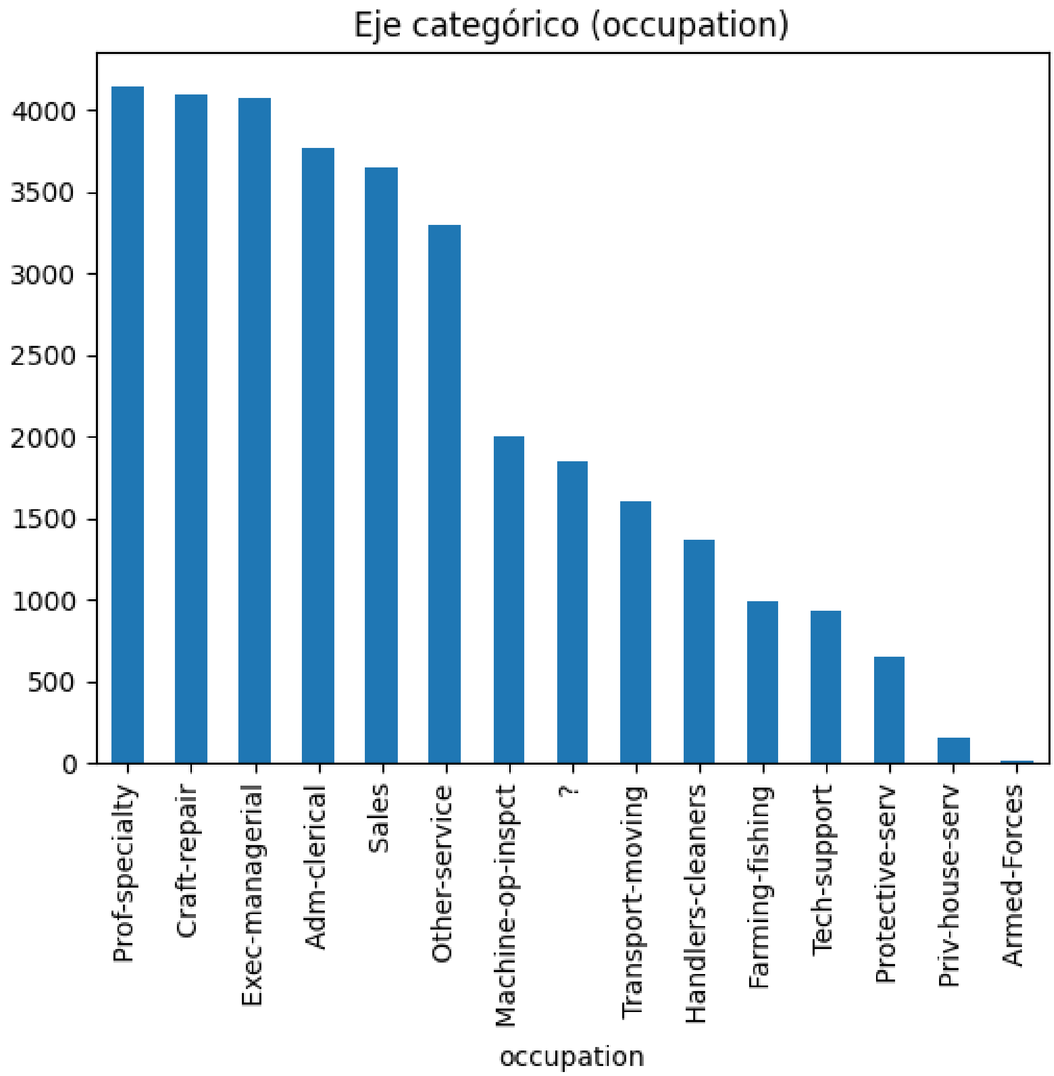

Eje categórico (workclass)
20000
15000
10000
5000
0
Private
Self-emp-not-inc
Local-gov
?
State-gov
Self-emp-inc
Federal-gov
Without-pay
Never-worked
workclass


Eje numérico (fnlwgt)
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
fnlwgt
fnlwgt

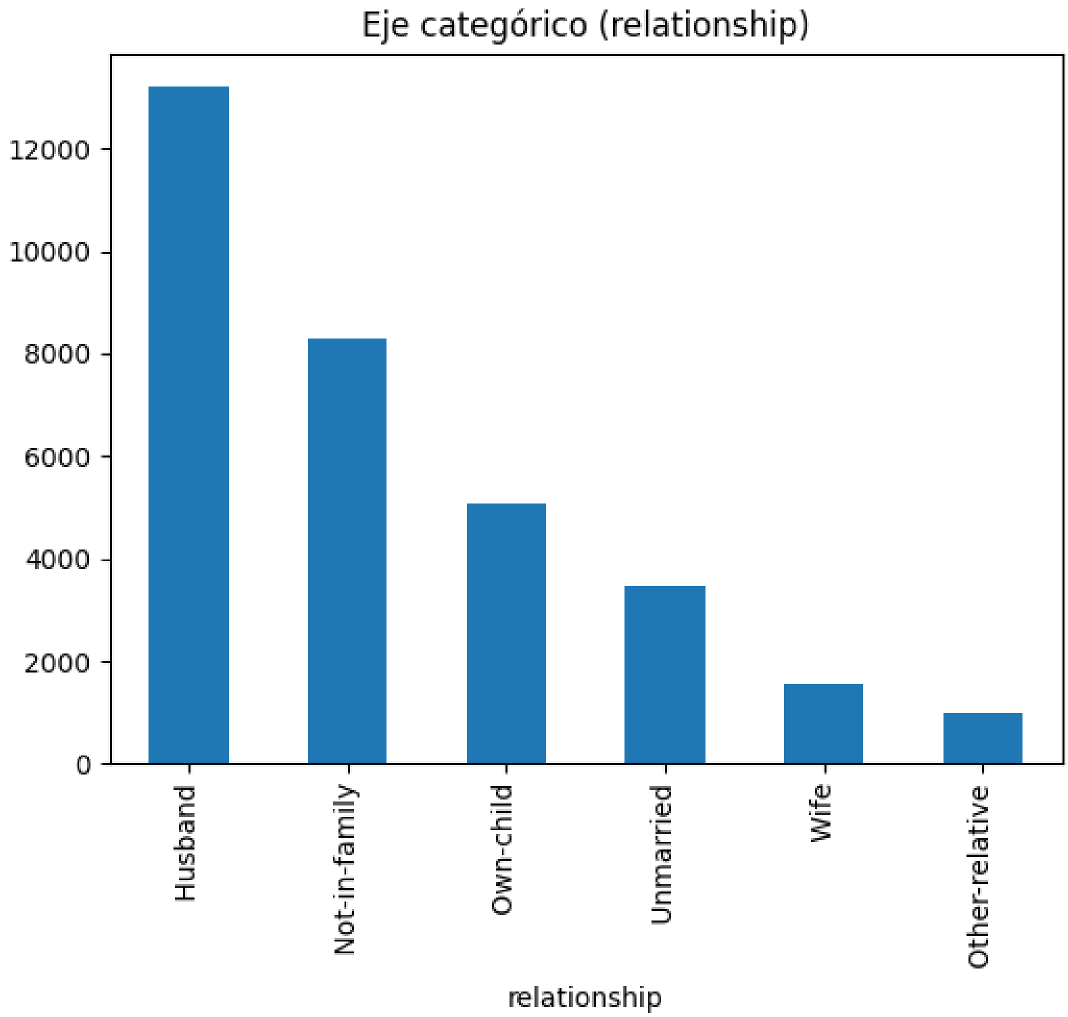

Eje categórico (education)
10000
8000
6000
4000
2000
0
HS-grad
Some-college
Bachelors
Masters
Assoc-voc
11th
Assoc-acdm
10th
7th-8th
Prof-school
9th
12th
Doctorate
5th-6th
1st-4th
Preschool
education


Eje numérico (education-num)
16
14
12
10
8
6
4
2
education-num
education-num

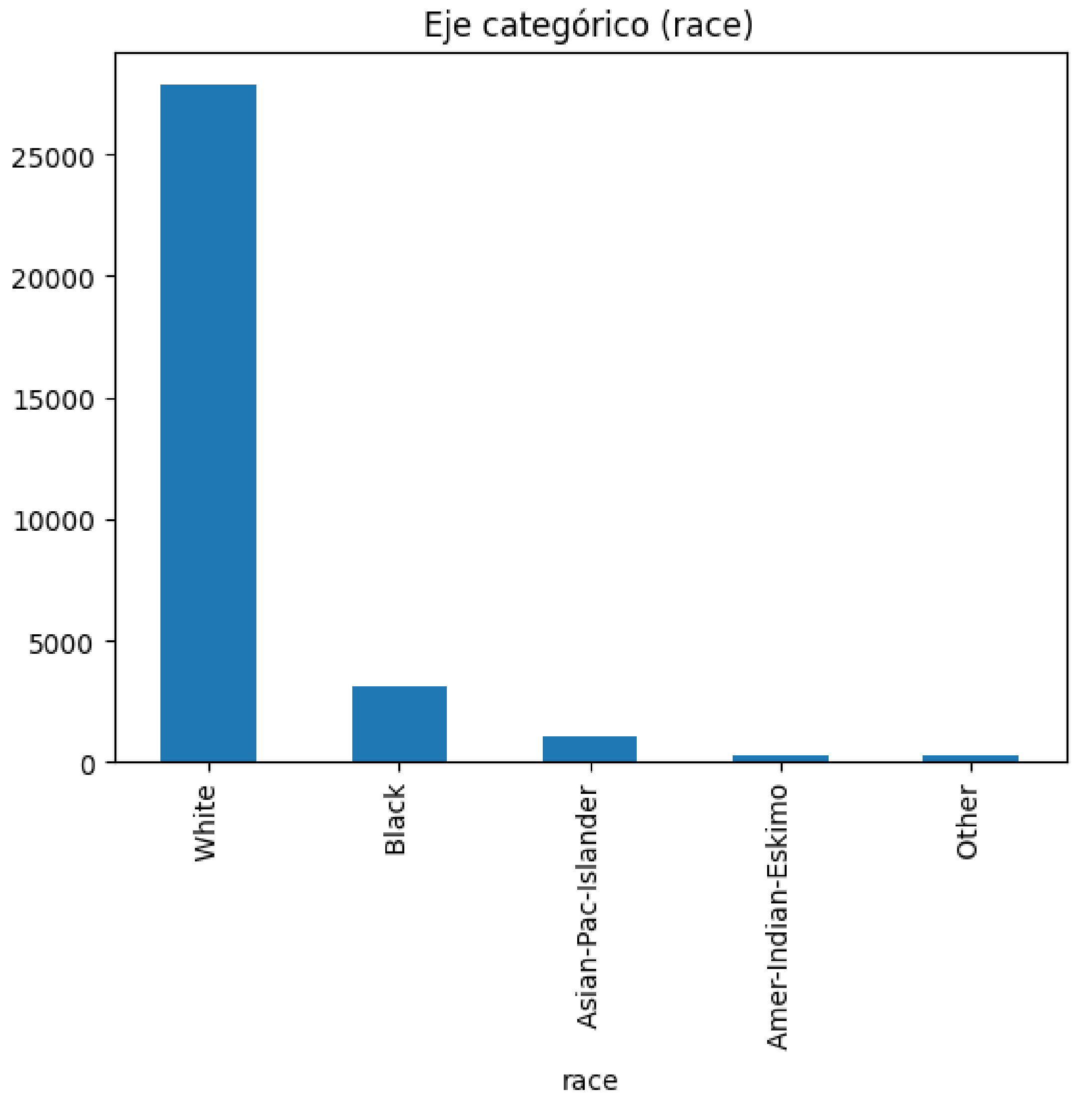

Eje categórico (marital-status)
14000
12000
10000
8000
6000
4000
2000
0
Married-civ-spouse
Never-married
Divorced
Separated
Widowed
Married-spouse-absent
Married-AF-spouse
marital-status

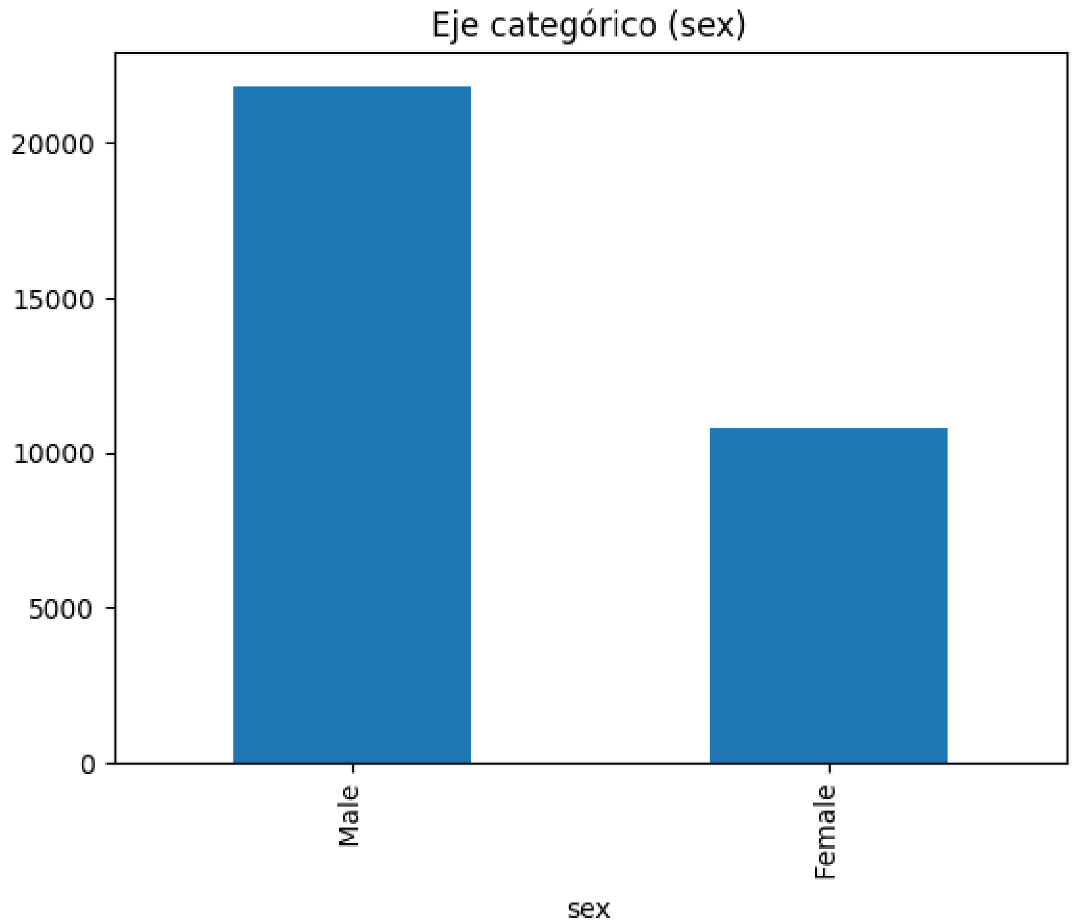


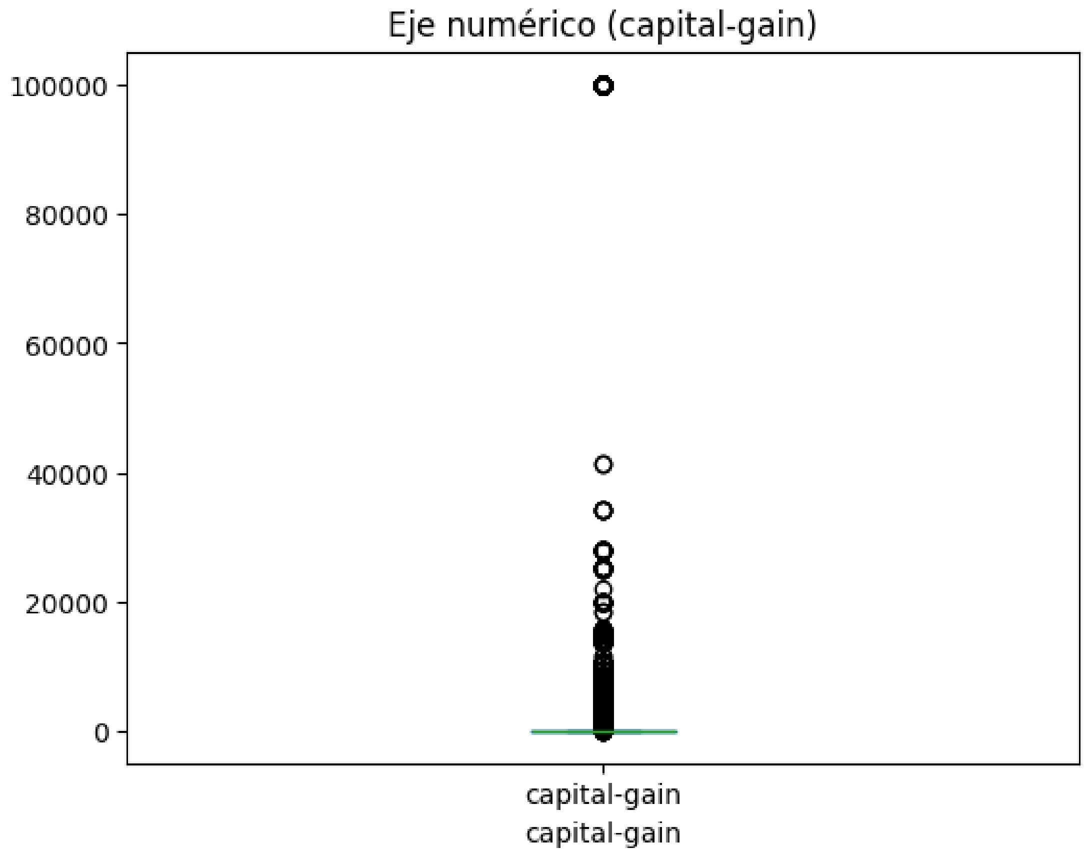
Eje categórico (occupation)
4000
3500
3000
2500
2000
1500
1000
500
0
Prof-specialty
Craft-repair
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
?
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces
occupation

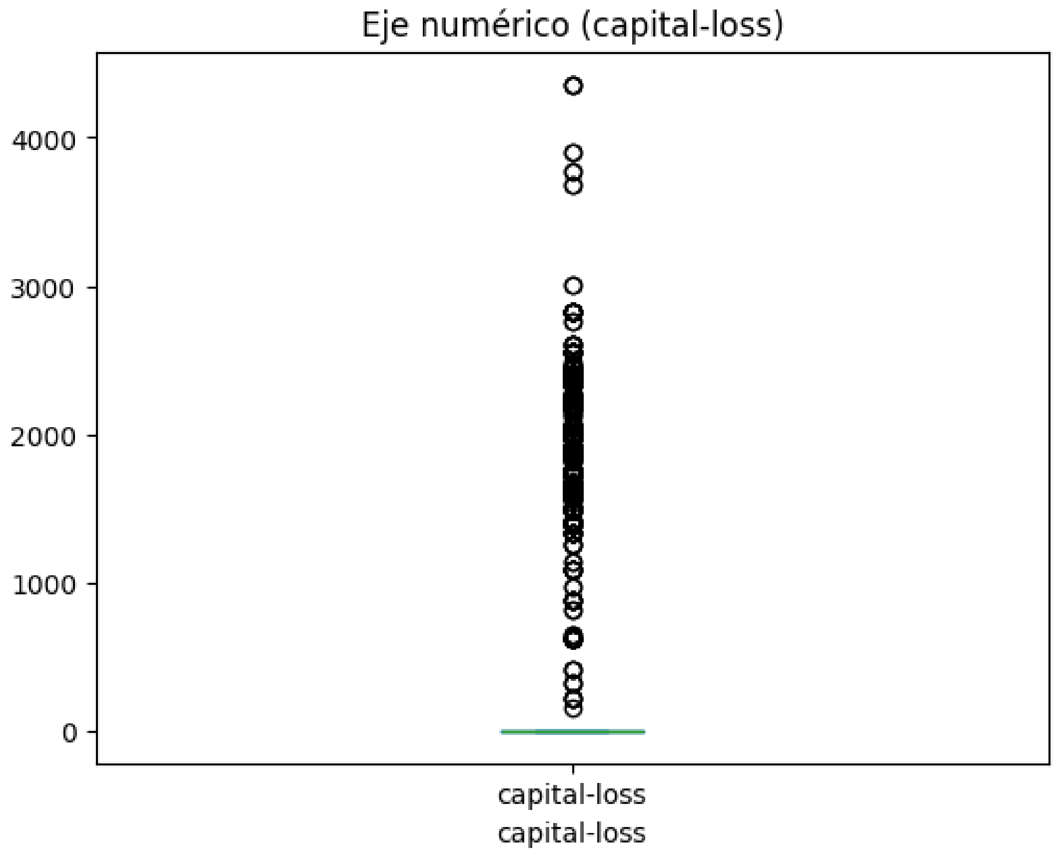


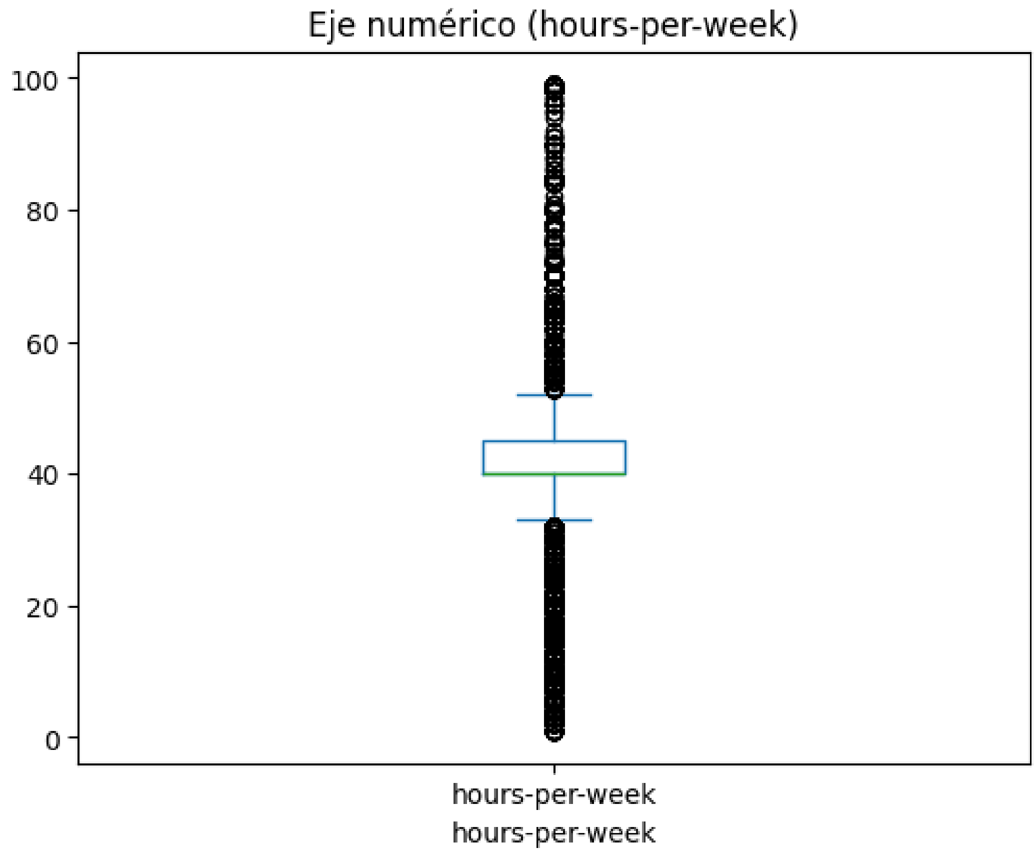
Eje categórico (relationship)
12000
10000
8000
6000
4000
2000
0
Husband
Not-in-family
Own-child
Unmarried
Wife
Other-relative
relationship

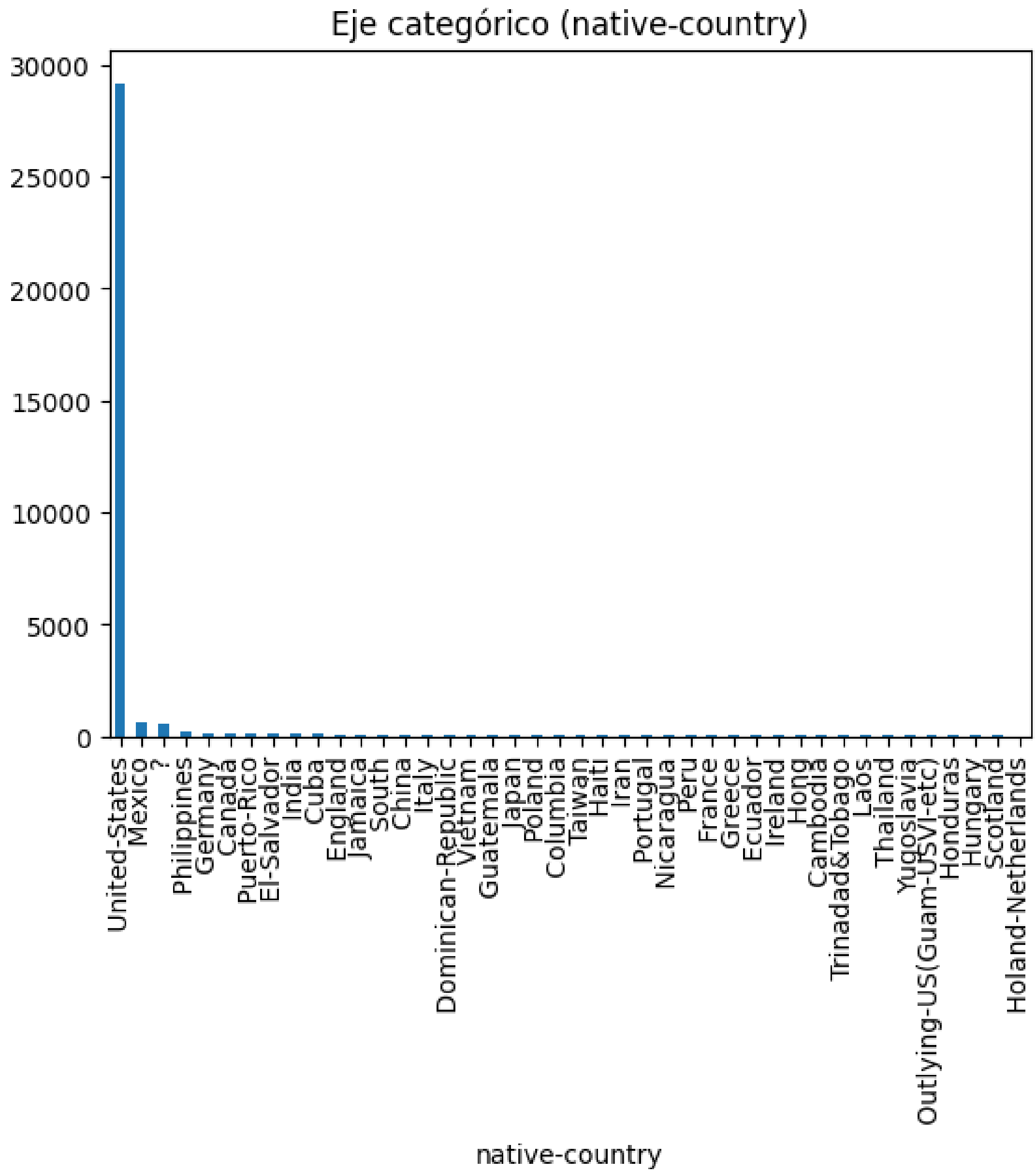

Eje categórico (race)
25000
20000
15000
10000
5000
0
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other
race

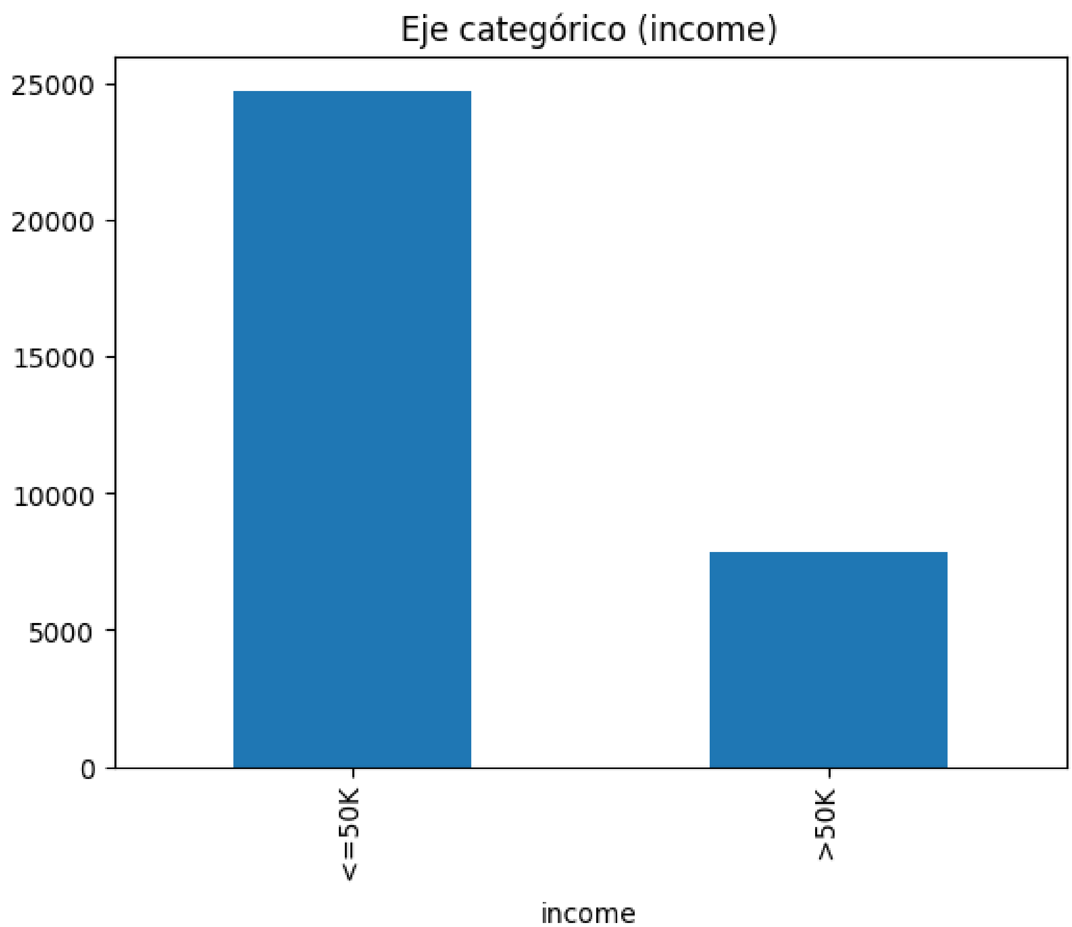

Eje categórico (sex)
20000
15000
10000
5000
0
Male
Female
sex


Eje numérico (capital-gain)
100000
80000
60000
40000
20000
0
capital-gain
capital-gain


Eje numérico (capital-loss)
4000
3000
2000
1000
0
capital-loss
capital-loss


Eje numérico (hours-per-week)
100
80
60
40
20
0
hours-per-week
hours-per-week


Eje categórico (native-country)
30000
25000
20000
15000
10000
5000
0
United-States
Mexico
?
Philippines
Germany
Canada
Puerto-Rico
El-Salvador
India
Cuba
England
Jamaica
South
China
Italy
Dominican-Republic
Vietnam
Guatemala
Japan
Poland
Columbia
Taiwan
Haiti
Iran
Portugal
Nicaragua
Peru
France
Greece
Ecuador
Ireland
Hong
Cambodia
Trinadad&Tobago
Laos
Thailand
Yugoslavia
Outlying-US(Guam-USVI-etc)
Honduras
Hungary
Scotland
Holand-Netherlands
native-country

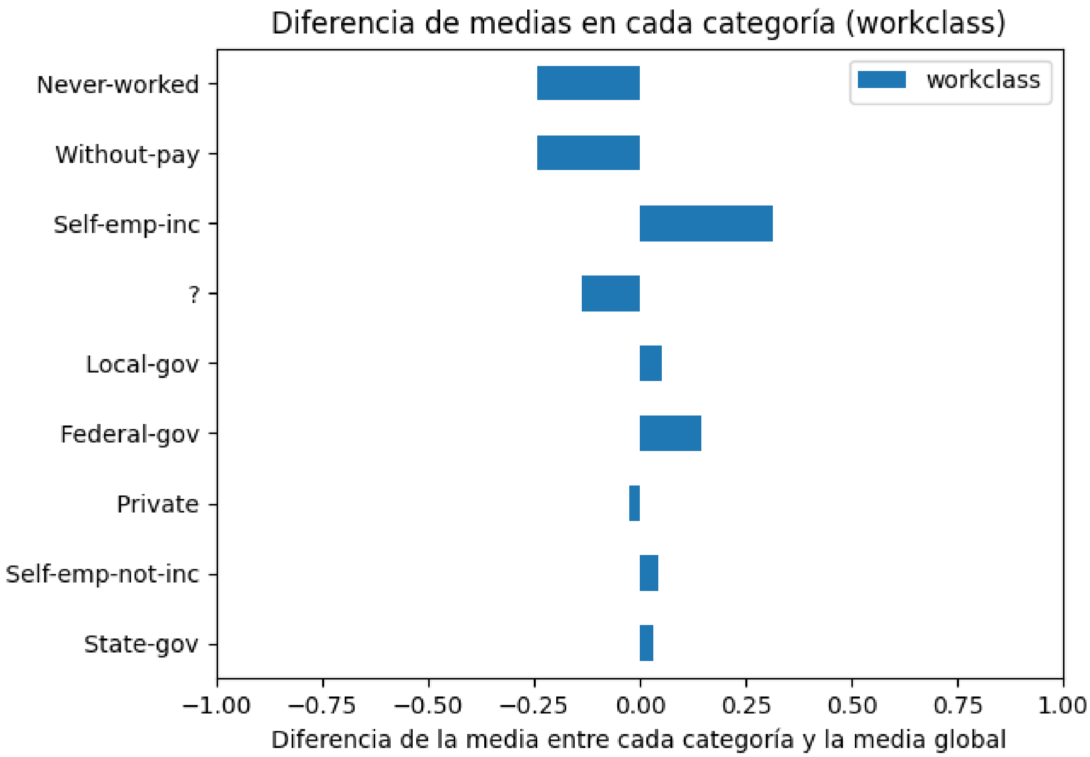


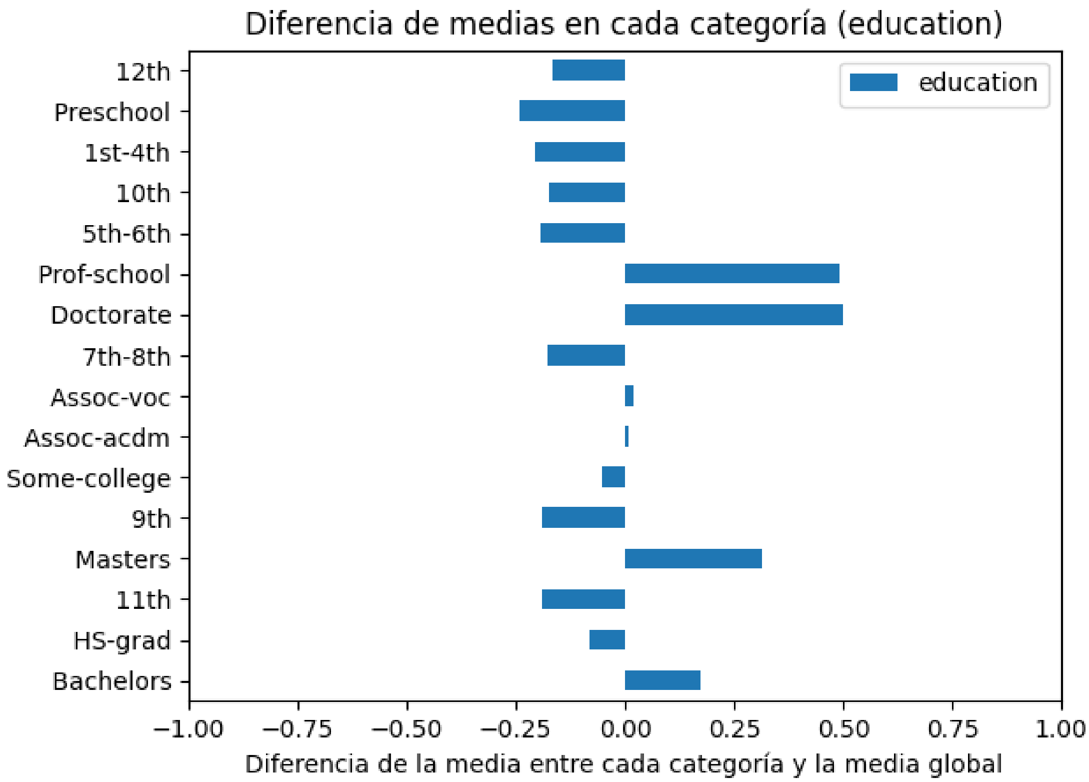
- **age**: En edad observamos puntos atípicos y podríamos estratificar la edad o tomarla como una variable continua o normalizada, por ejemplo, de la menor a la mayor edad o por segmentos de edades.
- **workclass**: En tipo de trabajo observamos que la mayoría son del sector privado y los demás se dividen en los puestos gubernamentales, auto-empleados y que no trabajan. Además hay una categoría donde están los desconocidos (`?`).
- **relationship**: En tipo de relación la mayoría es esposo y las demás pueden ser *dummies*.

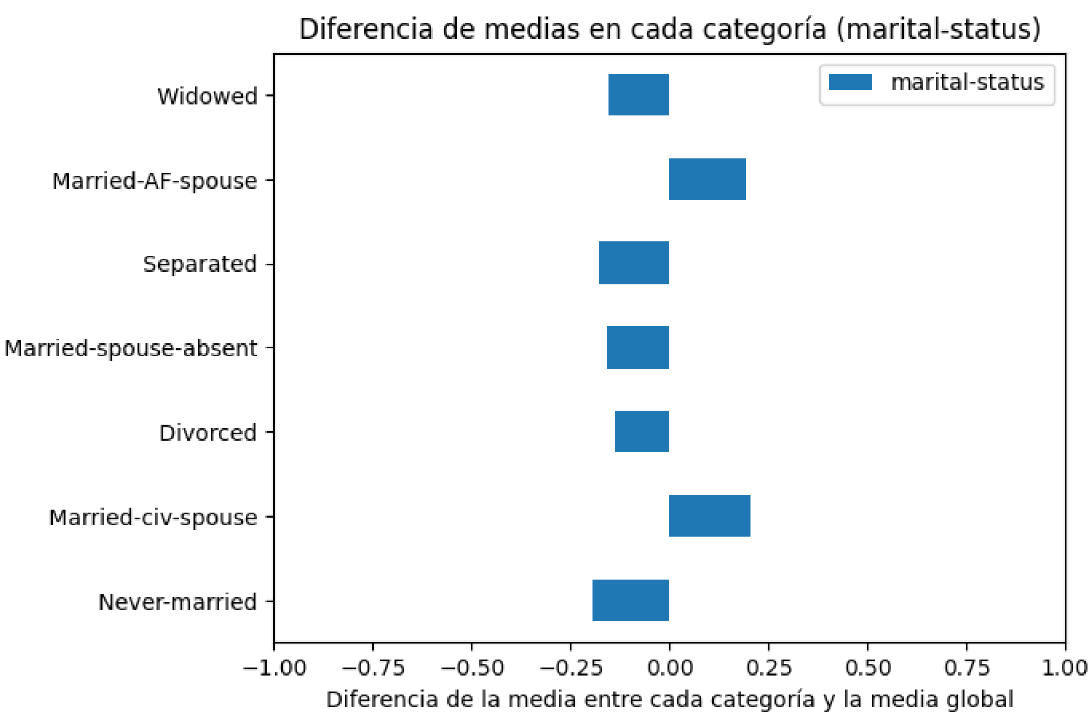
```
array([' <=50K', ' >50K'], dtype=object)
```

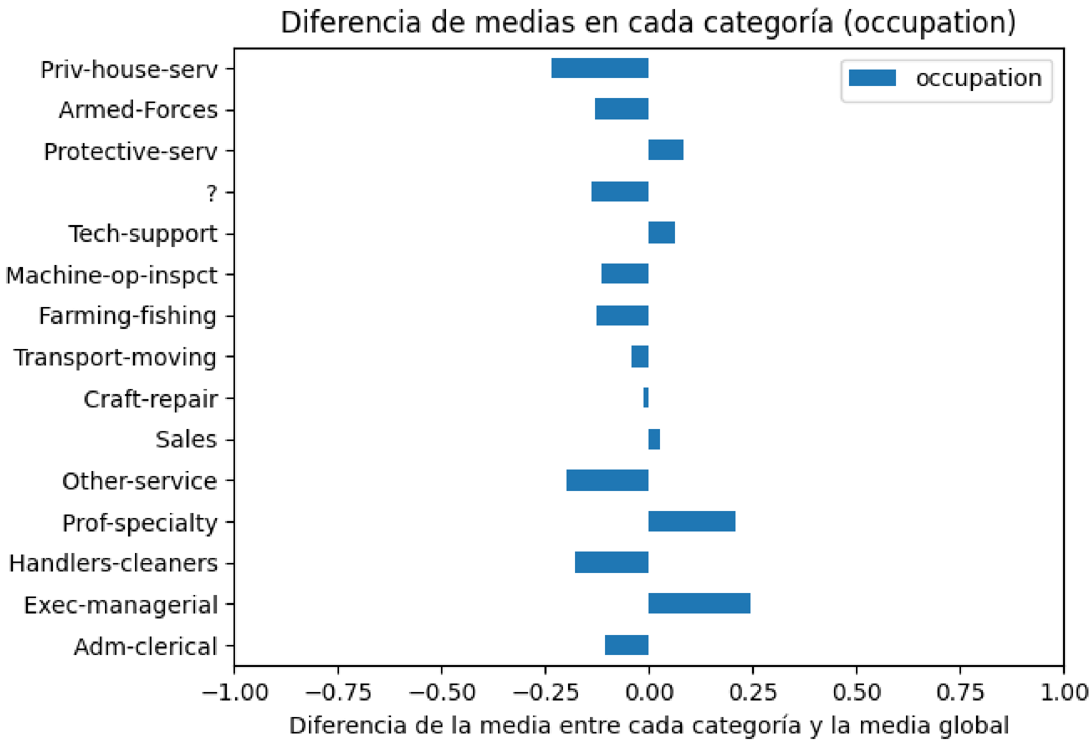
```
30366    0
28347    0
26379    0
5898     1
12674    0
Name: income, dtype: int64
```

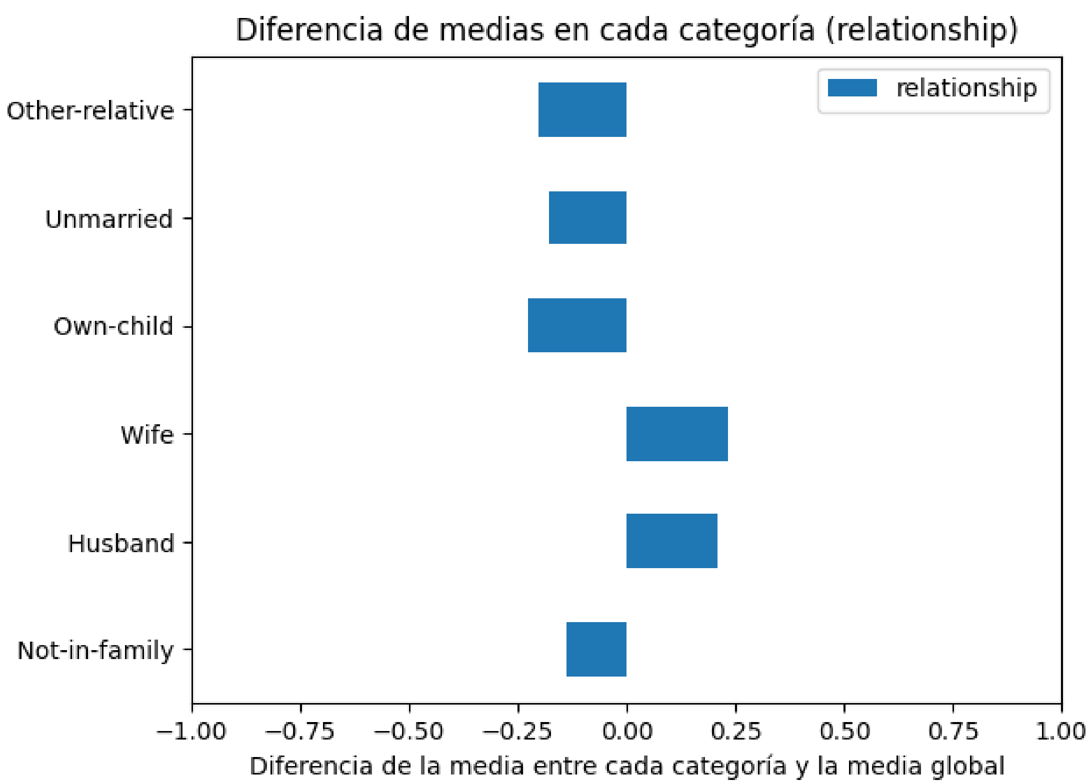
```
array([' State-gov', ' Self-emp-not-inc', ' Private', ' Federal-gov',
       ' Local-gov', ' ?', ' Self-emp-inc', ' Without-pay',
       ' Never-worked'], dtype=object)
```

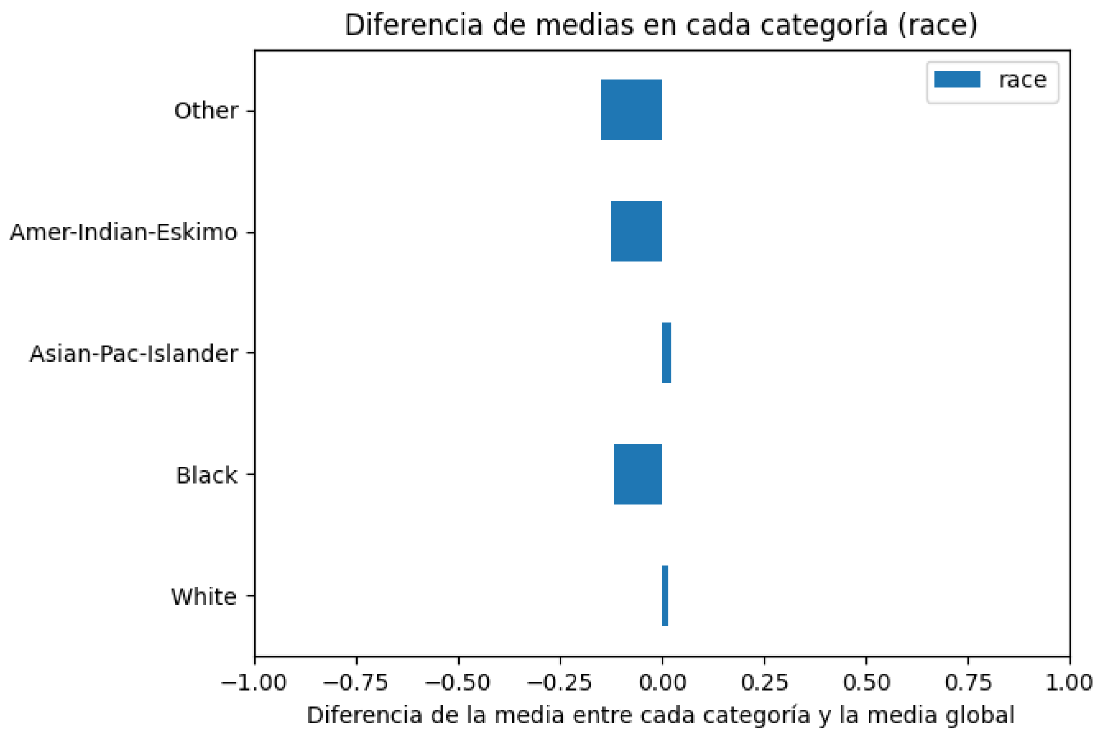



Diferencia de medias en cada categoría (education)
education
12th
Preschool
1st-4th
10th
5th-6th
Prof-school
Doctorate
7th-8th
Assoc-voc
Assoc-acdm
Some-college
9th
Masters
11th
HS-grad
Bachelors
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Diferencia de la media entre cada categoría y la media global

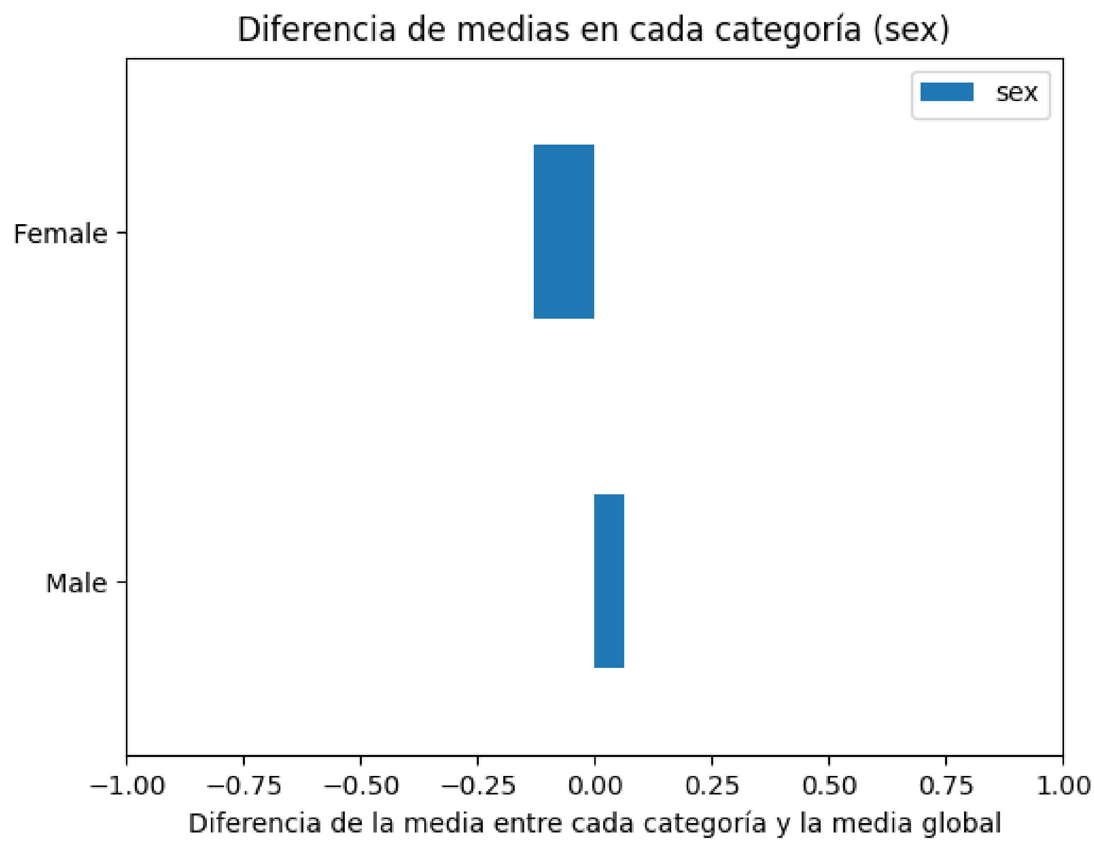

Diferencia de medias en cada categoría (marital-status)
marital-status
Widowed
Married-AF-spouse
Separated
Married-spouse-absent
Divorced
Married-civ-spouse
Never-married
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Diferencia de la media entre cada categoría y la media global


Diferencia de medias en cada categoría (occupation)
occupation
Priv-house-serv
Armed-Forces
Protective-serv
?
Tech-support
Machine-op-inspct
Farming-fishing
Transport-moving
Craft-repair
Sales
Other-service
Prof-specialty
Handlers-cleaners
Exec-managerial
Adm-clerical
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Diferencia de la media entre cada categoría y la media global


Diferencia de medias en cada categoría (relationship)
relationship
Other-relative
Unmarried
Own-child
Wife
Husband
Not-in-family
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Diferencia de la media entre cada categoría y la media global



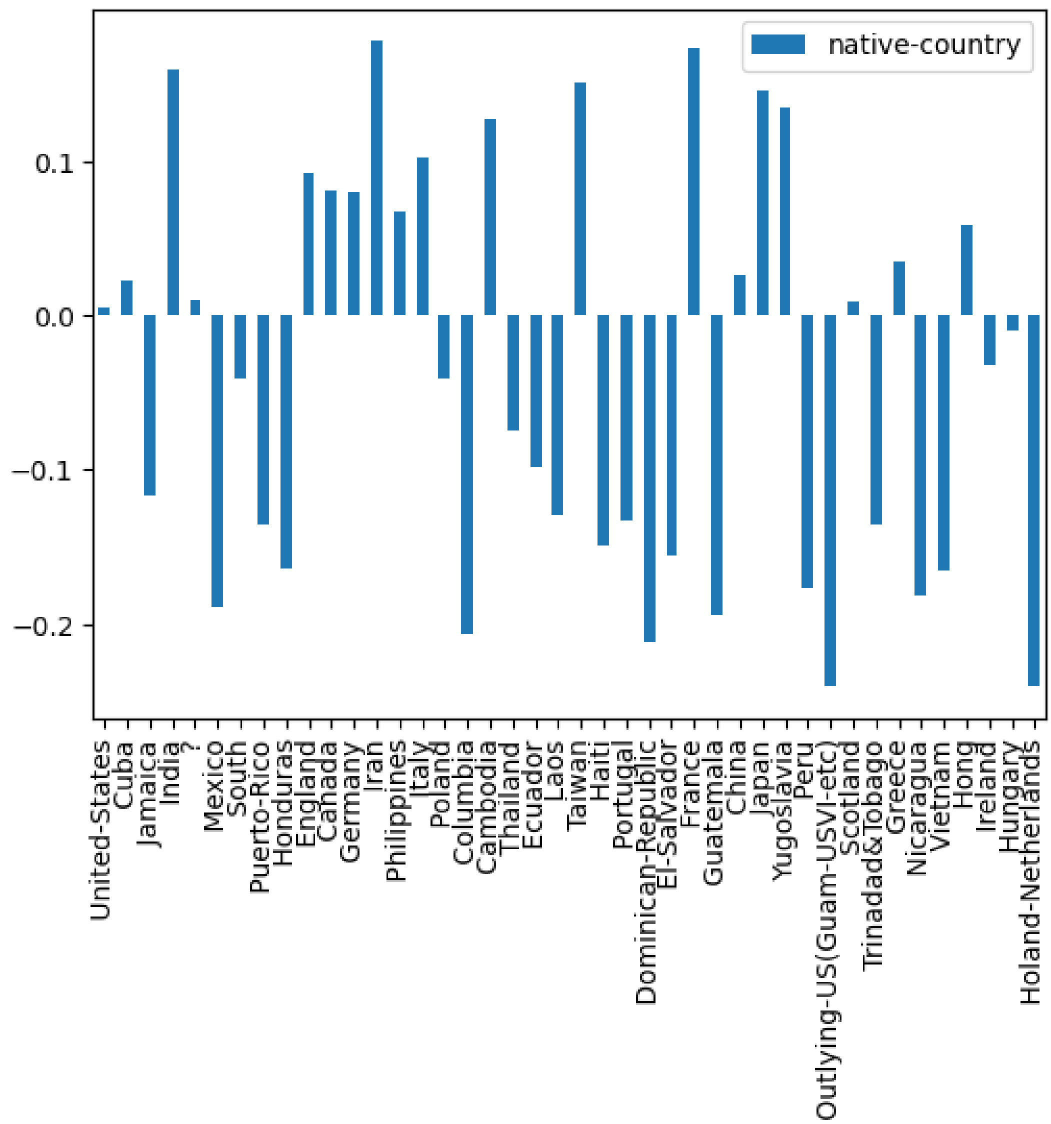
Diferencia de medias en cada categoría (race)
race
Other
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
White
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Diferencia de la media entre cada categoría y la media global



	native-country
Iran	0.177795
France	0.172984
India	0.15919
Taiwan	0.151347
Japan	0.146287
Yugoslavia	0.13419
Cambodia	0.127611
Italy	0.101656
England	0.092524
Canada	0.081504
Germany	0.080358
Philippines	0.067271
Hong	0.05919
Greece	0.035053
China	0.025857
Cuba	0.022348
?	0.009619
Scotland	0.00919
United-States	0.005025

```
Hungary                          -0.01004
Ireland                         -0.032476
South                            -0.04081
Poland                           -0.04081
Thailand                        -0.074143
Ecuador                         -0.097952
Jamaica                         -0.117353
Laos                            -0.129698
Portugal                        -0.132701
Trinadad&Tobago                 -0.135546
Puerto-Rico                     -0.135546
Haiti                             -0.1499
El-Salvador                     -0.155904
Honduras                        -0.163886
Vietnam                         -0.166183
Peru                            -0.176293
Nicaragua                       -0.181986
Mexico                          -0.189488
Guatemala                       -0.193935
Columbia                        -0.206911
Dominican-Republic              -0.212238
Outlying-US(Guam-USVI-etc)       -0.24081
Holand-Netherlands               -0.24081
```

```
<Axes: >
```



	native-country	clu
Iran	0.177795	0
France	0.172984	0
India	0.15919	0
Taiwan	0.151347	0
Japan	0.146287	0
Yugoslavia	0.13419	4
Cambodia	0.127611	4
Italy	0.101656	4
England	0.092524	7
Canada	0.081504	7
Germany	0.080358	7

```
Philippines                     0.067271    7
Hong                             0.05919    7
Greece                          0.035053    2
China                           0.025857    2
Cuba                            0.022348    2
?                               0.009619    2
Scotland                         0.00919    2
United-States                   0.005025    2
Hungary                         -0.01004    2
Ireland                        -0.032476    5
South                           -0.04081    5
Poland                          -0.04081    5
Thailand                       -0.074143    5
Ecuador                        -0.097952    1
Jamaica                        -0.117353    1
Laos                           -0.129698    1
Portugal                       -0.132701    1
Trinadad&Tobago                -0.135546    1
Puerto-Rico                    -0.135546    1
Haiti                            -0.1499    6
El-Salvador                    -0.155904    6
Honduras                       -0.163886    6
Vietnam                        -0.166183    6
Peru                           -0.176293    6
Nicaragua                      -0.181986    6
Mexico                         -0.189488    6
Guatemala                      -0.193935    6
Columbia                       -0.206911    3
Dominican-Republic             -0.212238    3
Outlying-US(Guam-USVI-etc)      -0.24081    3
Holand-Netherlands              -0.24081    3
```

```
0     1
1     3
2     4
3     5
4     3
5     8
6    10
dtype: int64
```

```
0    0
1    1
2    0
3    0
4    1
5    0
```

```
6    0
dtype: int64
```

3.1 Construcción de las varibles

```
           Categoría
0          State-gov
1   Self-emp-not-inc
2            Private
3        Federal-gov
4          Local-gov
5                  ?
6       Self-emp-inc
7        Without-pay
8       Never-worked

        Categoría
0       Bachelors
1         HS-grad
2            11th
3         Masters
4             9th
5    Some-college
6      Assoc-acdm
7       Assoc-voc
8         7th-8th
9       Doctorate
10    Prof-school
11        5th-6th
12           10th
13        1st-4th
14      Preschool
15           12th

                Categoría
0           Never-married
1      Married-civ-spouse
2                Divorced
3   Married-spouse-absent
4               Separated
5       Married-AF-spouse
6                 Widowed

             Categoría
0         Adm-clerical
1      Exec-managerial
2    Handlers-cleaners
```

```
3       Prof-specialty
4        Other-service
5                Sales
6          Craft-repair
7     Transport-moving
8      Farming-fishing
9    Machine-op-inspct
10        Tech-support
11                   ?
12     Protective-serv
13        Armed-Forces
14     Priv-house-serv
```

```
        Categoría
0    Not-in-family
1          Husband
2             Wife
3        Own-child
4        Unmarried
5   Other-relative
```

```
            Categoría
0               White
1               Black
2  Asian-Pac-Islander
3  Amer-Indian-Eskimo
4               Other
```

```
  Categoría
0      Male
1    Female
```

```
                    Categoría
0               United-States
1                        Cuba
2                     Jamaica
3                       India
4                           ?
5                      Mexico
6                       South
7                 Puerto-Rico
8                    Honduras
9                     England
10                     Canada
11                    Germany
12                       Iran
13                Philippines
```

```
14                       Italy
15                      Poland
16                    Columbia
17                    Cambodia
18                    Thailand
19                     Ecuador
20                        Laos
21                      Taiwan
22                       Haiti
23                    Portugal
24          Dominican-Republic
25                 El-Salvador
26                      France
27                   Guatemala
28                       China
29                       Japan
30                  Yugoslavia
31                        Peru
32  Outlying-US(Guam-USVI-etc)
33                    Scotland
34             Trinadad&Tobago
35                      Greece
36                   Nicaragua
37                     Vietnam
38                        Hong
39                     Ireland
40                     Hungary
41          Holand-Netherlands

    x1   x2   x3   x4   x5   x6   x7   x8   x9  x10  …  x18  x19  x20  x21  \
0  0.0  0.0  0.0  1.0  0.0  0.0  1.0  0.0  0.0  0.0  …  0.0  0.0  0.0  0.0
1  0.0  0.0  0.0  1.0  0.0  0.0  1.0  0.0  0.0  1.0  …  0.0  0.0  0.0  0.0
2  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  …  0.0  0.0  0.0  0.0
3  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0  1.0  …  0.0  0.0  0.0  0.0
4  0.0  0.0  0.0  0.0  0.0  0.0  1.0  0.0  0.0  1.0  …  0.0  0.0  0.0  0.0

    x22       x23   x24  x25  x26   x27
0  39.0   77516.0  13.0  1.0  0.0  40.0
1  50.0   83311.0  13.0  0.0  0.0  32.5
2  38.0  215646.0   9.0  0.0  0.0  40.0
3  53.0  234721.0   7.0  0.0  0.0  40.0
4  28.0  338409.0  13.0  0.0  0.0  40.0

[5 rows x 27 columns]
```

4 Fase 3 - Modelos de Clasificación

```
income
0    0.759175
1    0.240825
Name: proportion, dtype: float64

income
0    0.759251
1    0.240749
Name: proportion, dtype: float64
```

4.1 Reporte

Exactitud Proporción de predicciones correctas sobre el total de casos.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precisión (Precision) Qué proporción de las predicciones positivas fueron correctas.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Sensibilidad (Recall o TPR) Qué proporción de los positivos reales fueron correctamente identificados.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Especificidad (TNR) Qué proporción de los negativos reales fueron correctamente identificados.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

F1-score Media armónica entre precisión y recall.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

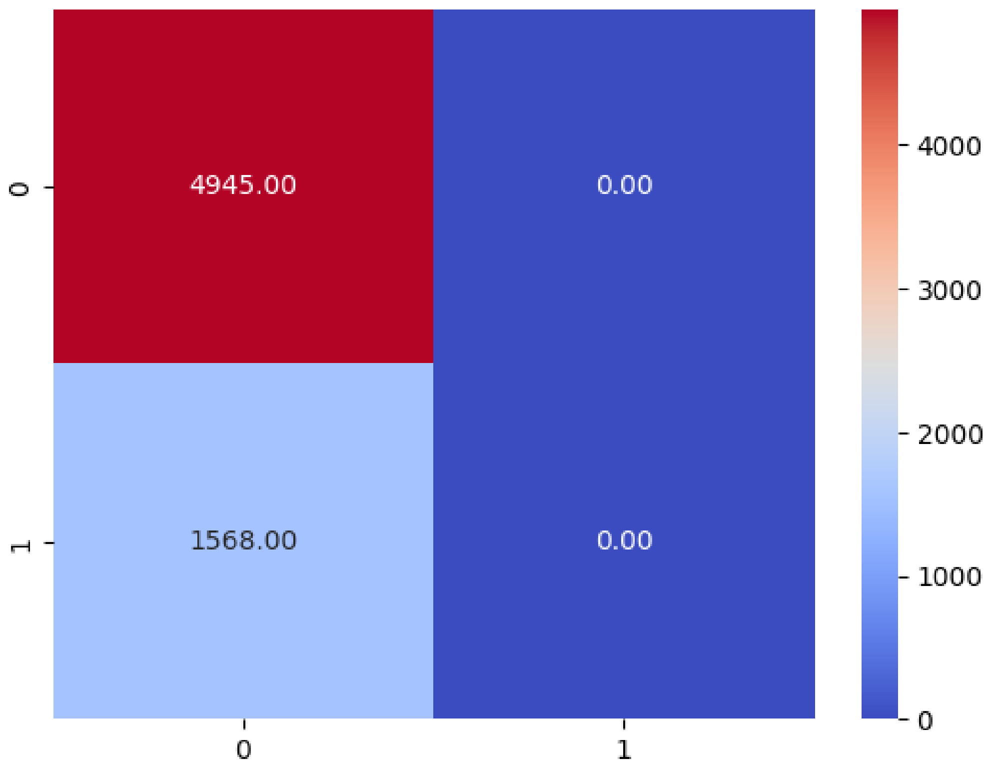
[[TN, FP], [FN, TP]]

- TN: True Negatives (verdaderos negativos)
- FP: False Positives (falsos positivos)
- FN: False Negatives (falsos negativos)
- TP: True Positives (verdaderos positivos)

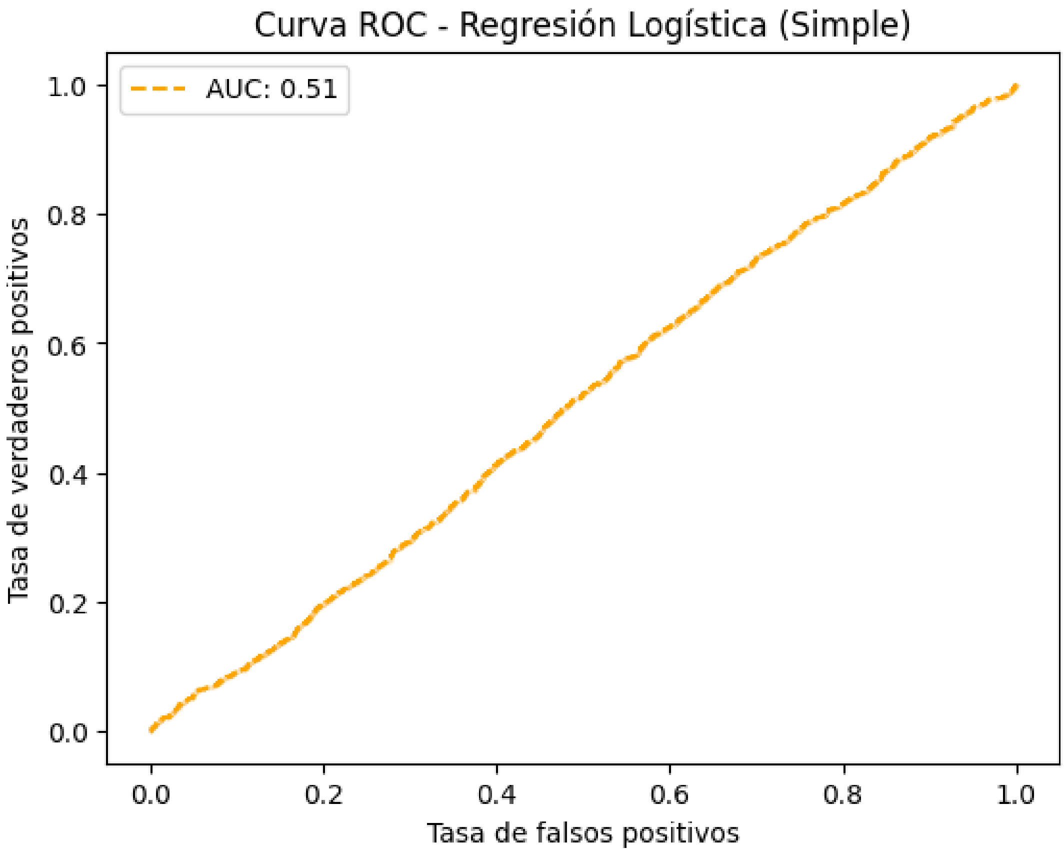
4.2 Regresión Logística

```
/Library/Frameworks/Python.framework/Versions/3.12/lib/python3.12/site-
packages/sklearn/linear_model/_sag.py:349: ConvergenceWarning: The max_iter was
reached which means the coef_ did not converge
  warnings.warn(
```

```
0.7592507293106096

<Axes: >
```




```
/var/folders/zr/py0pd6bs6gnfzg9ljbgr9v0c0000gn/T/ipykernel_981/610393776.py:5:
RuntimeWarning: invalid value encountered in scalar divide
  precision = TP / (TP + FP)

                  Valor
Exactitud      0.759251
Presición           NaN
Sensibilidad   0.000000
Especificidad  1.000000
F1                  NaN
```

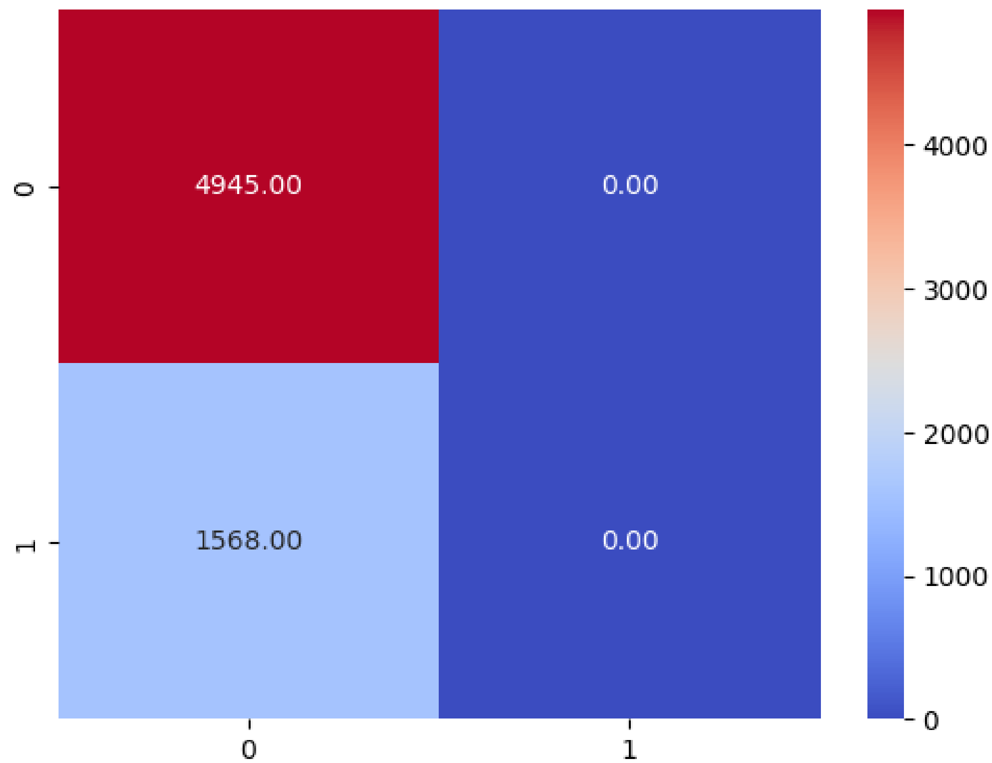


4.3 Regresión Logística Lasso

```
/Library/Frameworks/Python.framework/Versions/3.12/lib/python3.12/site-
packages/sklearn/linear_model/_sag.py:349: ConvergenceWarning: The max_iter was
reached which means the coef_ did not converge
  warnings.warn(
```

```
0.7592507293106096
```

```
<Axes: >
```



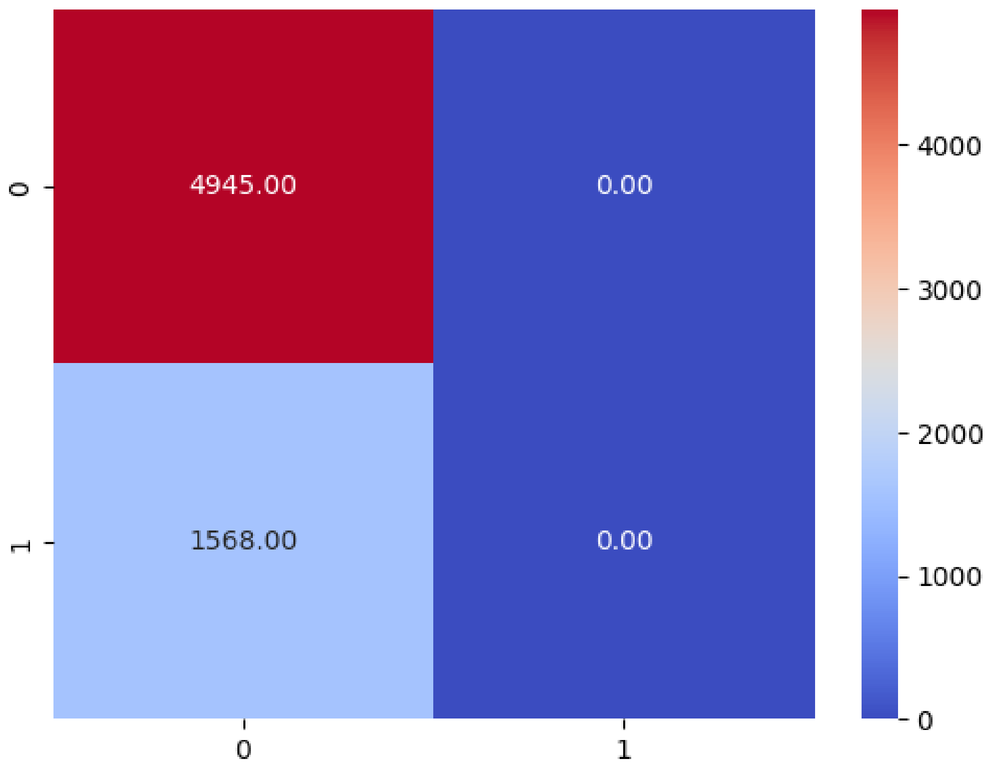

```
/var/folders/zr/py0pd6bs6gnfzg9ljbgr9v0c0000gn/T/ipykernel_981/610393776.py:5:
RuntimeWarning: invalid value encountered in scalar divide
  precision = TP / (TP + FP)

                 Valor
Exactitud      0.759251
Presición           NaN
Sensibilidad   0.000000
Especificidad  1.000000
F1                  NaN
```

4.4 Regresión Logística Ridge

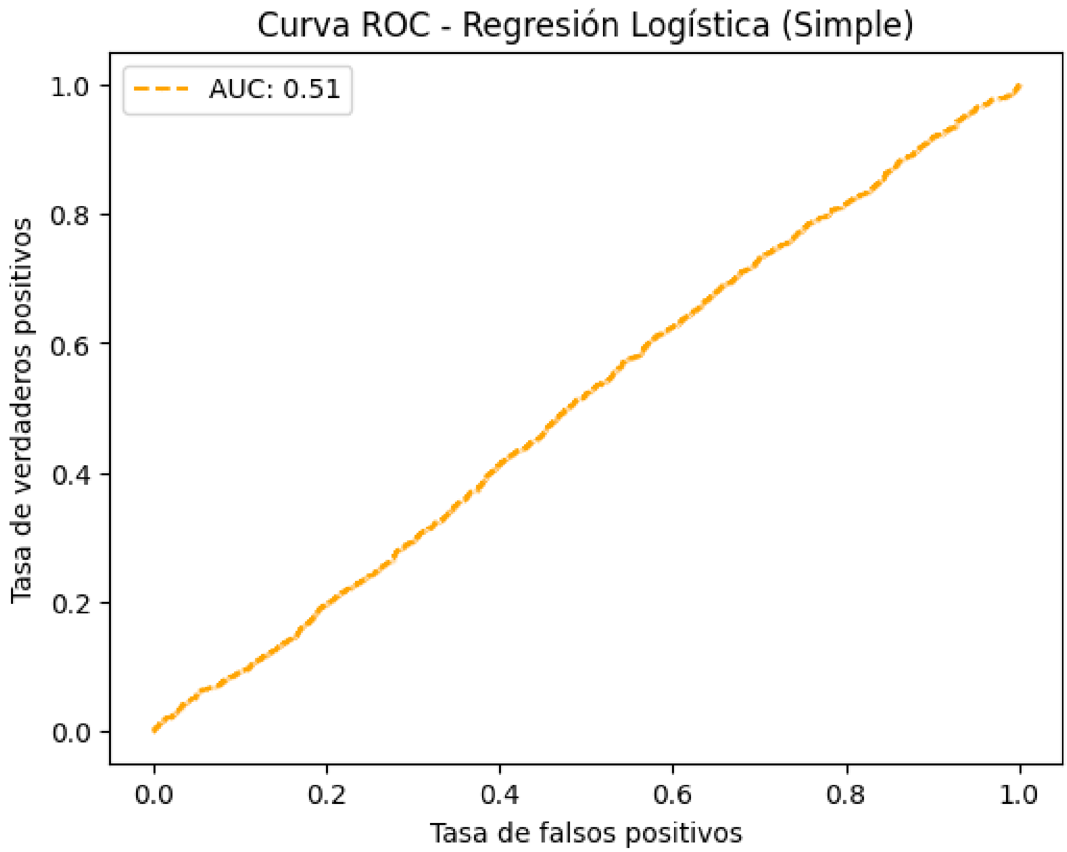
```
/Library/Frameworks/Python.framework/Versions/3.12/lib/python3.12/site-
packages/sklearn/linear_model/_sag.py:349: ConvergenceWarning: The max_iter was
reached which means the coef_ did not converge
  warnings.warn(
```

```
0.7592507293106096
```

```
<Axes: >
```



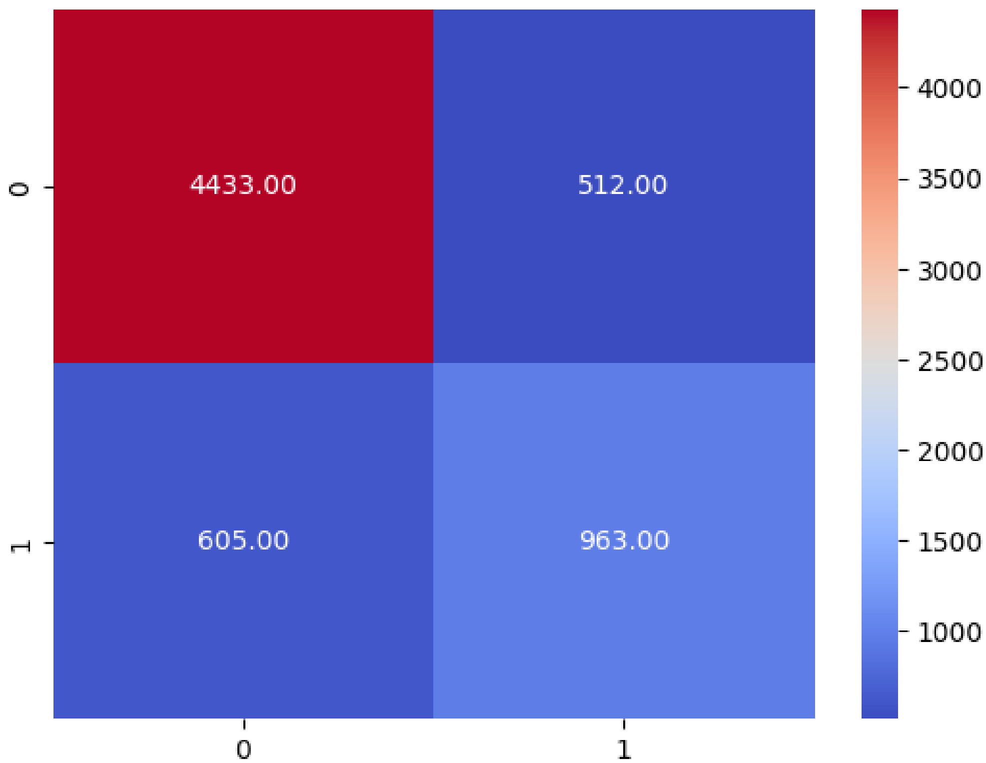

```
/var/folders/zr/py0pd6bs6gnfzg9ljbgr9v0c0000gn/T/ipykernel_981/610393776.py:5:
RuntimeWarning: invalid value encountered in scalar divide
  precision = TP / (TP + FP)

                  Valor
Exactitud      0.759251
Presición           NaN
Sensibilidad   0.000000
Especificidad  1.000000
F1                  NaN
```



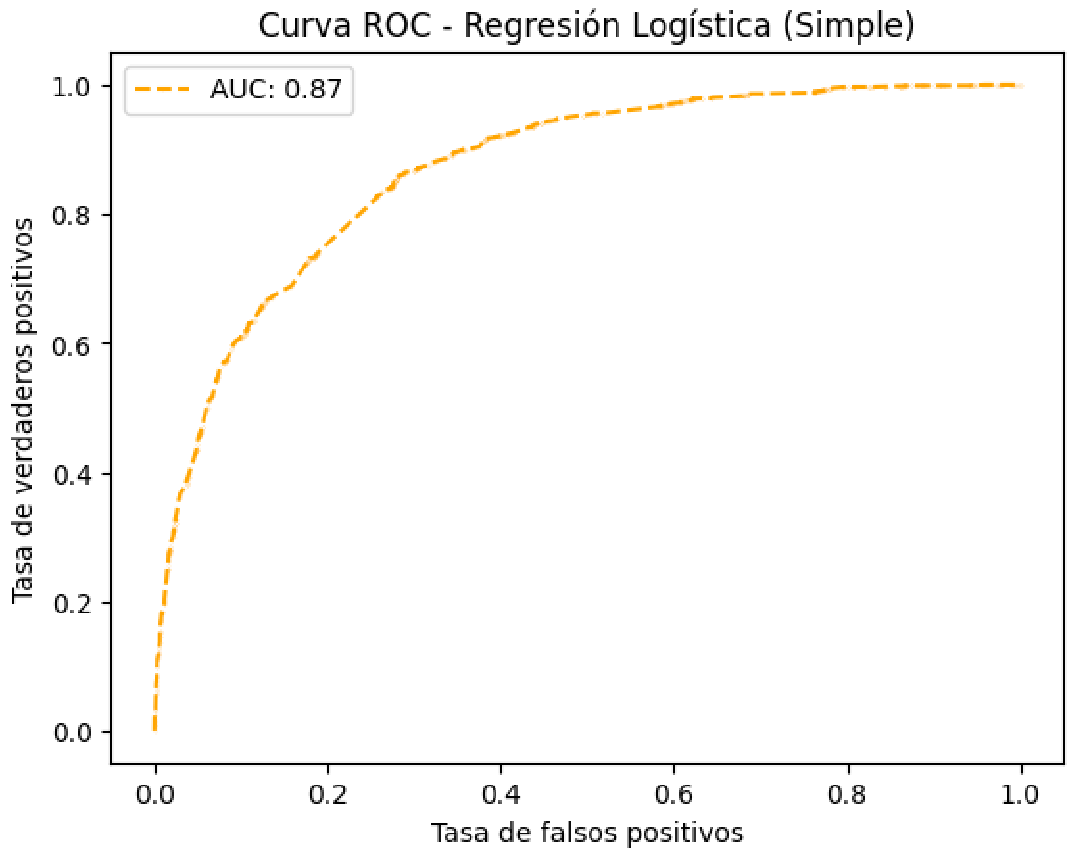
4.5 Naive Bayes

```
0.8284968524489482
```

```
<Axes: >
```



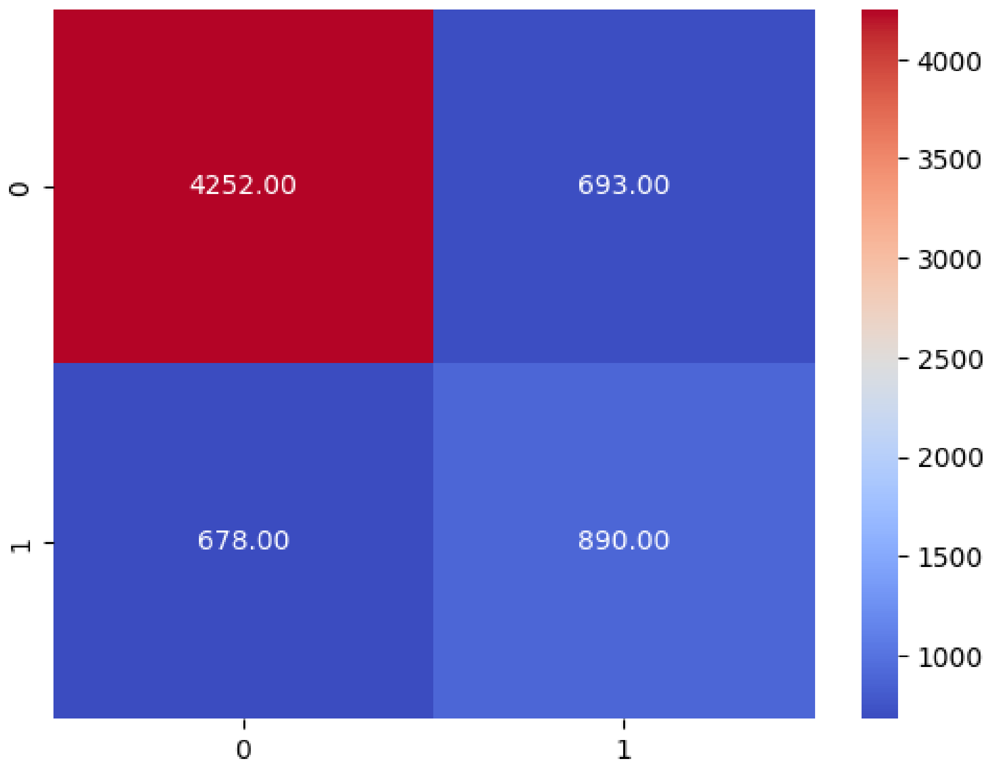
	Valor
Exactitud	0.828497
Presición	0.652881
Sensibilidad	0.614158
Especificidad	0.896461
F1	0.632928



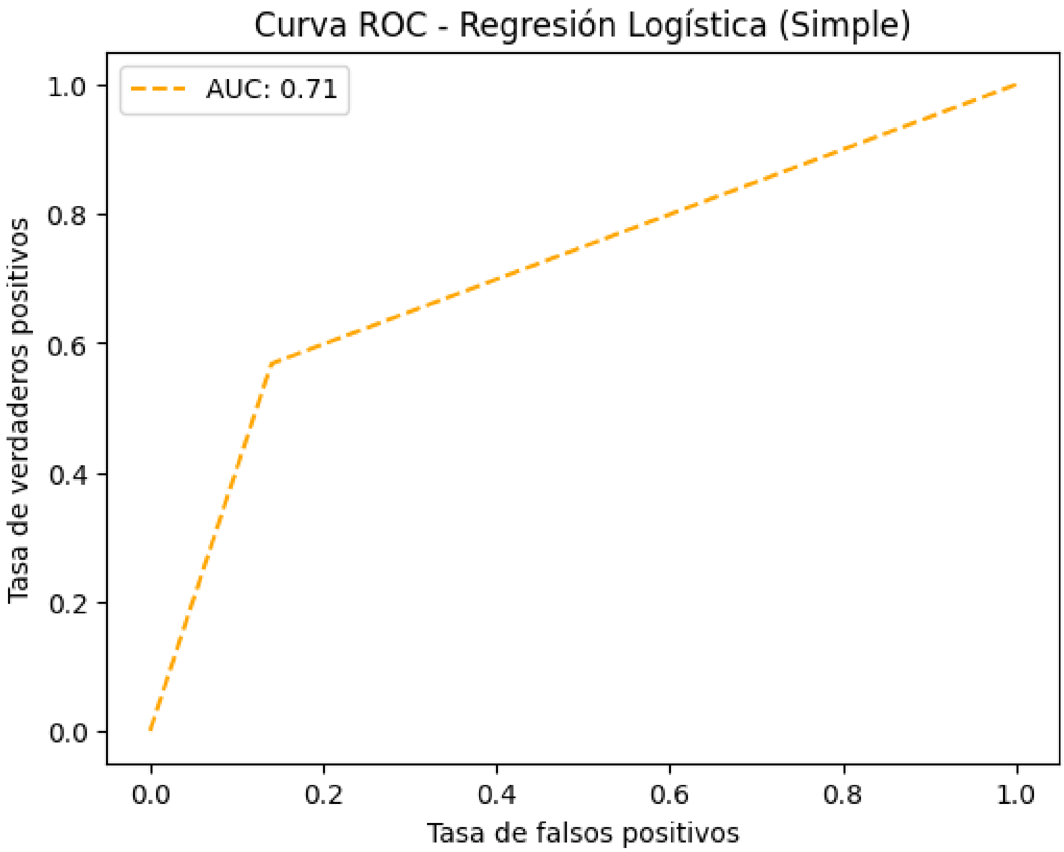
4.6 Árboles de Decisión

```
0.7894979272224781
```

```
<Axes: >
```



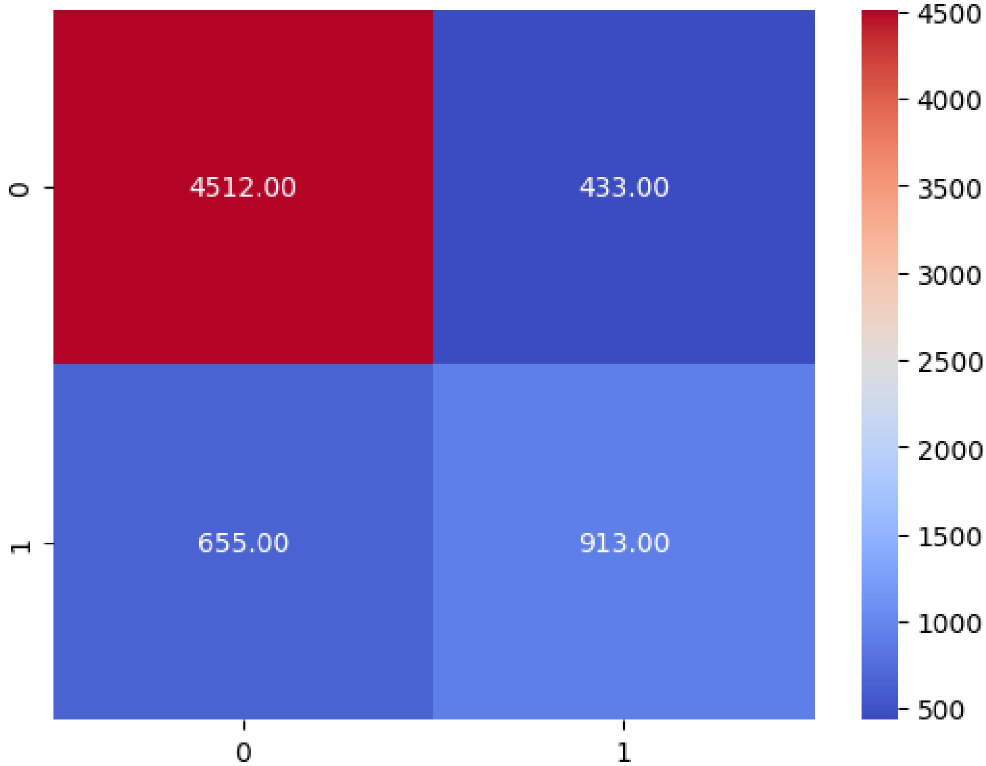
	Valor
Exactitud	0.789498
Presición	0.562224
Sensibilidad	0.567602
Especificidad	0.859858
F1	0.564900



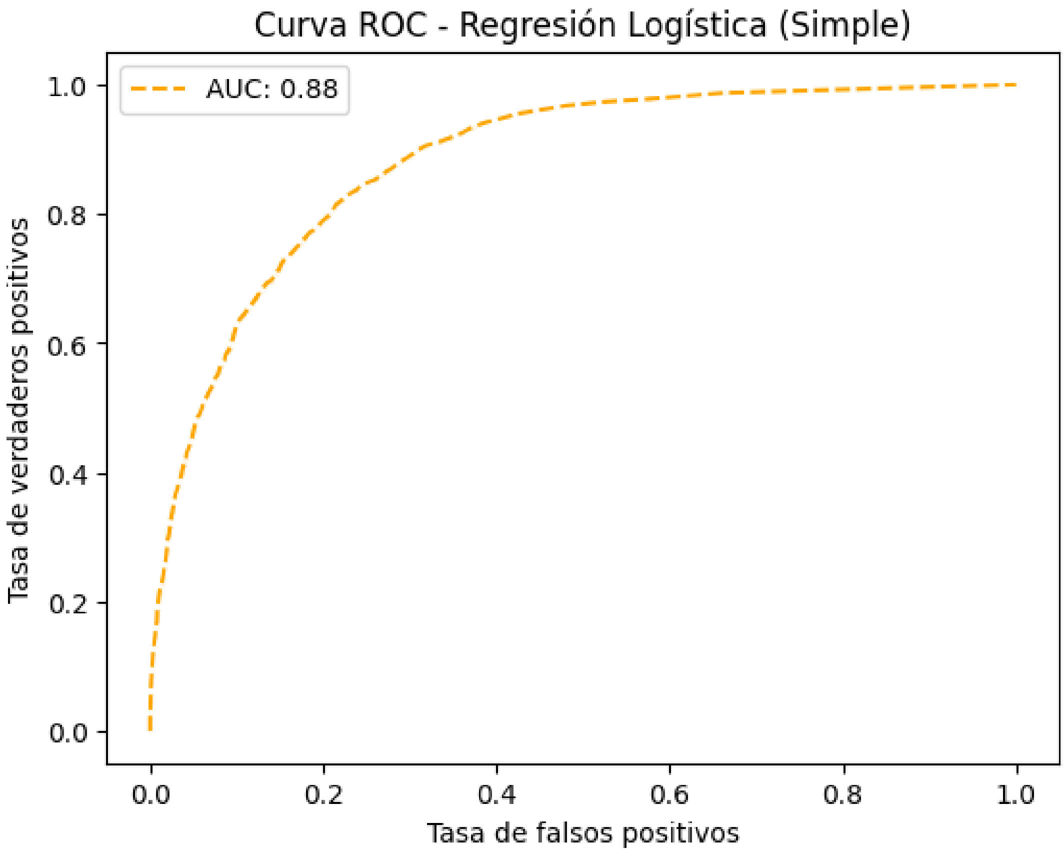
4.7 Bósques Aleatorios

```
0.8329494856440964
```

```
<Axes: >
```



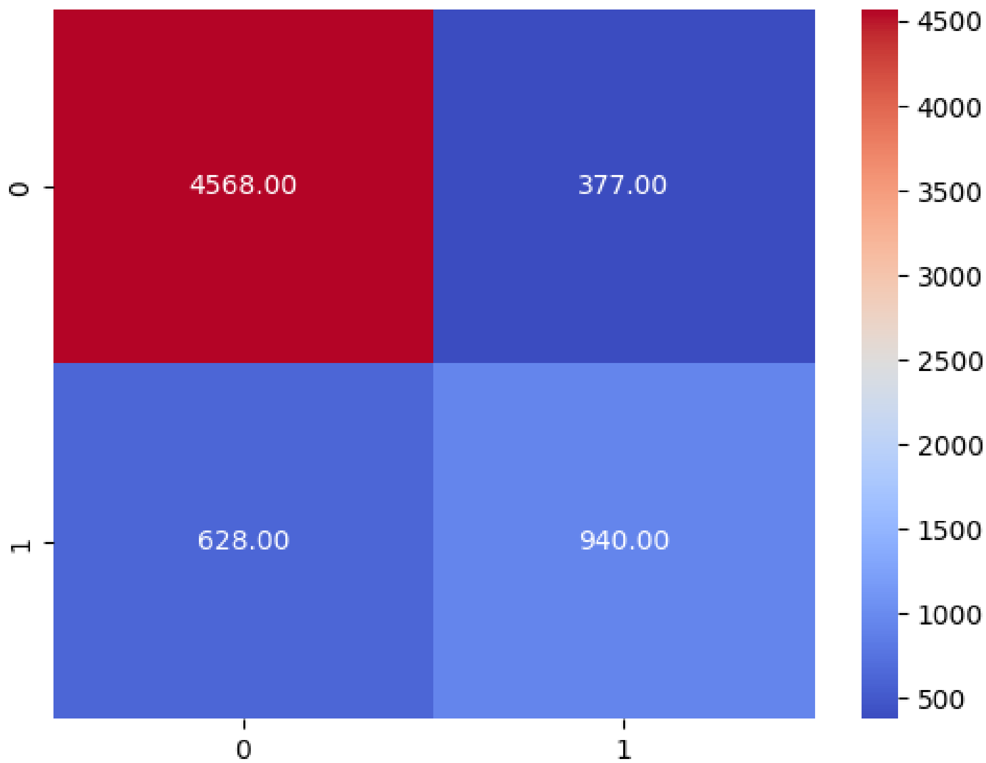
	Valor
Exactitud	0.832949
Presición	0.678306
Sensibilidad	0.582270
Especificidad	0.912437
F1	0.626630



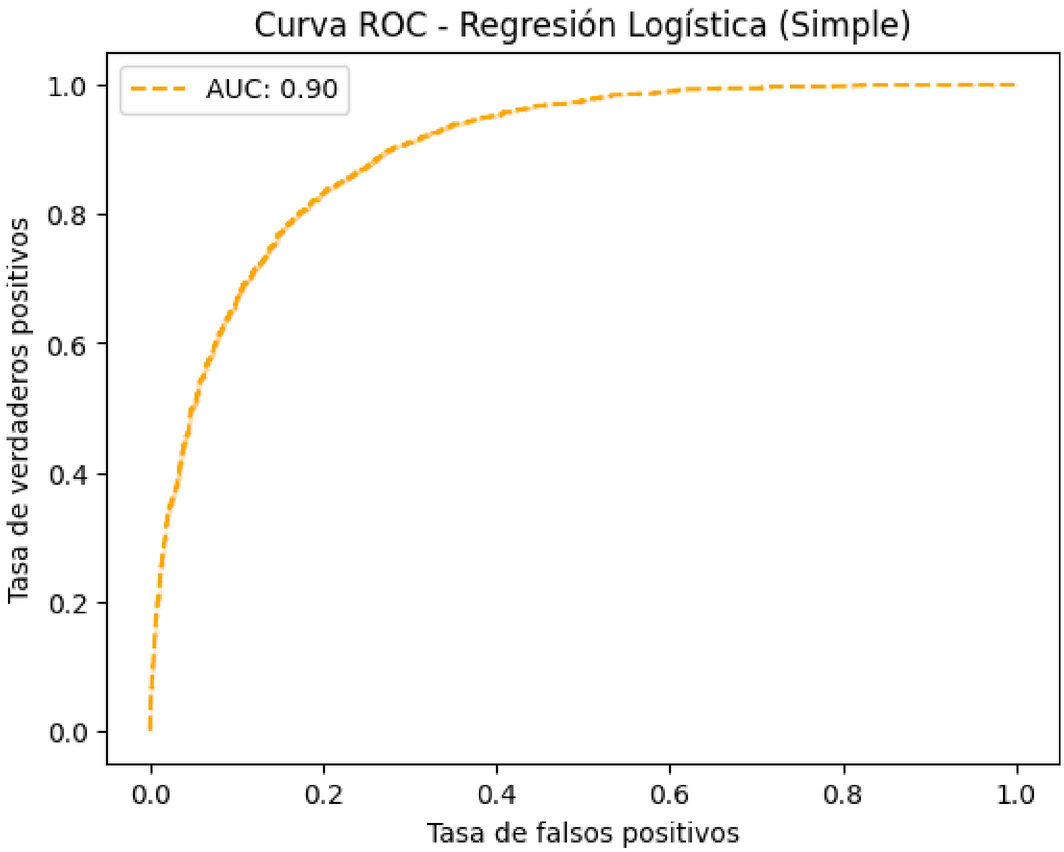
4.8 XGBoost

```
0.8456932289267619
```

```
<Axes: >
```



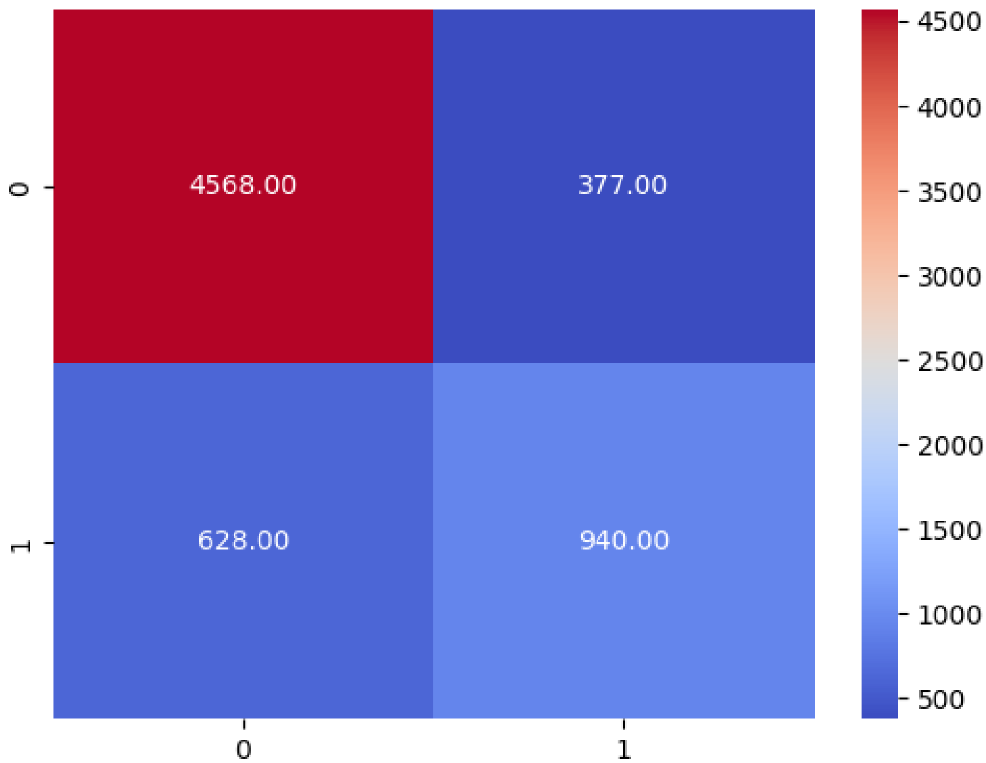
	Valor
Exactitud	0.845693
Presición	0.713743
Sensibilidad	0.599490
Especificidad	0.923761
F1	0.651646



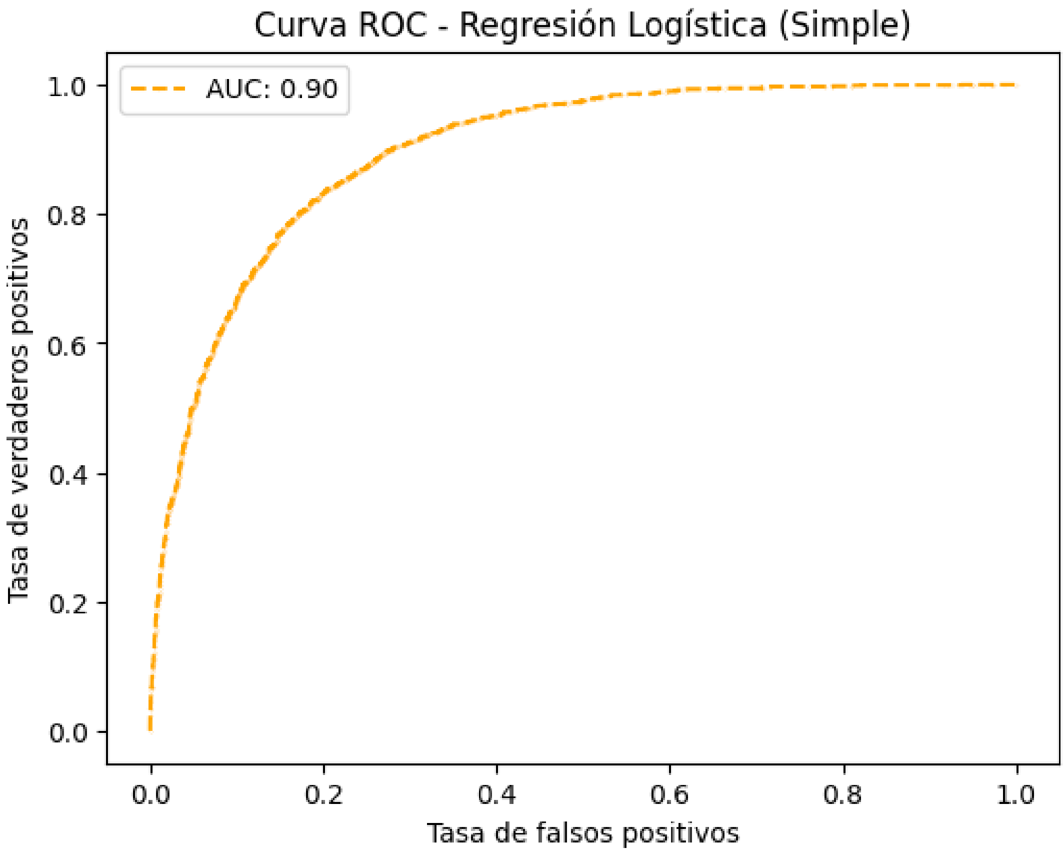
4.9 SVC

```
0.8456932289267619
```

```
<Axes: >
```



	Valor
Exactitud	0.845693
Presición	0.713743
Sensibilidad	0.599490
Especificidad	0.923761
F1	0.651646



5 Fase 4 - Ajuste del Modelo por Validación Cruzada

Hola