Nombre : Josue Alejandro Sauca Pucha

Fecha : 31-05-2023

Primer paso del ciclo de vida de datos

Como primer punto para el ciclo de vida de los datos se necesita la recopilacion de datos, esto se lo va a realizar al momento de recolectar todos los datos de diversas fuentes para tener una idea del panorama completo, es decir aqui se tiene una idea del contexto de lo que se va a hacer, en este caso se recopilo informacion de una base da datos libre del ecuador y se va a recopilar los datos de personas desaparecidas del año 2023

In [204]:
```
import pandas as pd
import numpy as np


#Leemos el dataset con la funcion read_csv
datos = pd.read_csv("mdg_personasdesaparecidas_pm_2023_enero_marzo.csv",sep=';
                    na_values = 'SIN_DATO')
```

In [205]:
```
#Vemos si el dataset se almaceno en la variable
datos
```

Out[205]:

	Provincia	Latitud	Longitud	Edad Aprox.	Sexo	Motivo Desaparición	Motivo Desaparición Obs
0	PICHINCHA	-0,2188216	-78,5135489	17	HOMBRE	NaN	Nal
1	PICHINCHA	-0,26909023	-78,54001523	17	MUJER	PROBLEMAS FAMILIARES	PROBLEMAS SENTIMENTALES
2	SANTO DOMINGO DE LOS TSACHILAS	0,0091672	-79,391605	39	HOMBRE	NaN	Nal
3	PICHINCHA	-0,17504166	-78,47478184	14	MUJER	PROBLEMAS SOCIALES	INFLUENCIA DE AMISTADES
4	ESMERALDAS	0,9873857	-79,65649069	28	HOMBRE	NaN	Nal
...	...	...	...	...	...	...	.
1887	SANTO DOMINGO DE LOS TSACHILAS	-0,2477616	-79,1485374	14	MUJER	NaN	Nal
1888	AZUAY	-3,38988084	-79,08271046	15	MUJER	NaN	Nal
1889	CHIMBORAZO	-1,908452	-78,641068	22	HOMBRE	PROBLEMAS SOCIALES	INFLUENCIA DE AMISTADES
1890	GUAYAS	-2,0721549	-79,9345141	32	HOMBRE	PROBLEMAS FAMILIARES	FAMILIA DISFUNSIONA
1891	PICHINCHA	-0,30864288	-78,54081268	38	HOMBRE	PROBLEMAS SOCIALES	INFLUENCIA DE AMISTADES

1892 rows × 10 columns

In [206]:
```
#Vemos los datos nulos del dataset
datos.isnull().sum()
```

Out[206]:
```
Provincia                  0
Latitud                    0
Longitud                   0
Edad Aprox.                0
Sexo                       0
Motivo Desaparición      286
Motivo Desaparición Obs.   286
Fecha Desaparición         0
Situación Actual           0
Fecha Localización       286
dtype: int64
```

Segundo paso del ciclo de vida de datos

Como segundo paso para la preparacion de datos se realiza la limpieza de datos, transformacion y reestructuracion de datos que se va a utilizar, en el caso que se esta presentando se va realizar un ingreso de datos en los valores nulos que existen dentro del dataset con el fin de que se pueda trabajar en el mismo, en el ejemplo se presenta un lleno a

In [207]:
```
#Rellanamos la columna Motivo Desaparicion con la media de dicha columna, para
#realizar la limpieza de datos, ya que en dicha columna se va a trabajar
datos['Motivo Desaparición'].fillna(datos['Motivo Desaparición'].mode()[0],inpla
```

In [208]:
```
#Como se puede observar dichas columnas ya se encuentra con los datos llenos
datos
```

Out[208]:

	Provincia	Latitud	Longitud	Edad Aprox.	Sexo	Motivo Desaparición	Motiv Desaparició Obs
0	PICHINCHA	-0,2188216	-78,5135489	17	HOMBRE	PROBLEMAS FAMILIARES	Nal
1	PICHINCHA	-0,26909023	-78,54001523	17	MUJER	PROBLEMAS FAMILIARES	PROBLEMAS SENTIMENTALES
2	SANTO DOMINGO DE LOS TSACHILAS	0,0091672	-79,391605	39	HOMBRE	PROBLEMAS FAMILIARES	Nal
3	PICHINCHA	-0,17504166	-78,47478184	14	MUJER	PROBLEMAS SOCIALES	INFLUENCIA DI AMISTADES
4	ESMERALDAS	0,9873857	-79,65649069	28	HOMBRE	PROBLEMAS FAMILIARES	Nal
...	...	...	...	...	...	...	.
1887	SANTO DOMINGO DE LOS TSACHILAS	-0,2477616	-79,1485374	14	MUJER	PROBLEMAS FAMILIARES	Nal
1888	AZUAY	-3,38988084	-79,08271046	15	MUJER	PROBLEMAS FAMILIARES	Nal
1889	CHIMBORAZO	-1,908452	-78,641068	22	HOMBRE	PROBLEMAS SOCIALES	INFLUENCIA DI AMISTADES
1890	GUAYAS	-2,0721549	-79,9345141	32	HOMBRE	PROBLEMAS FAMILIARES	FAMILIA DISFUNSIONA
1891	PICHINCHA	-0,30864288	-78,54081268	38	HOMBRE	PROBLEMAS SOCIALES	INFLUENCIA DI AMISTADES

1892 rows × 10 columns

Tercer paso del ciclo de vida de datos

Luego de realizar la preparacion de datos se continua con la eleccion de datos , para esto se va a eliminar las columnas que nos se requieran trabajar dentro del ejemplo presentado, ya que existe una condicional que nos limita a trabajar con una parte del dataset

a) Elimine las variables Latitud,Longitud,Motivo Desaparción Obs.,Fecha Desaparición,Fecha Localización

In [209]:
```
datos = datos.drop(columns=['Latitud', 'Longitud','Motivo Desaparción Obs.','F
```

In [210]:
```
#Imprimimos nuevamente el dataset con las columnas eliminadas
datos
```

Out[210]:

	Provincia	Edad Aprox.	Sexo	Motivo Desaparción	Situación Actual
0	PICHINCHA	17	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
1	PICHINCHA	17	MUJER	PROBLEMAS FAMILIARES	ENCONTRADO
2	SANTO DOMINGO DE LOS TSACHILAS	39	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
3	PICHINCHA	14	MUJER	PROBLEMAS SOCIALES	ENCONTRADO
4	ESMERALDAS	28	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
...	...	...	...	...	...
1887	SANTO DOMINGO DE LOS TSACHILAS	14	MUJER	PROBLEMAS FAMILIARES	DESAPARECIDO
1888	AZUAY	15	MUJER	PROBLEMAS FAMILIARES	DESAPARECIDO
1889	CHIMBORAZO	22	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO
1890	GUAYAS	32	HOMBRE	PROBLEMAS FAMILIARES	ENCONTRADO
1891	PICHINCHA	38	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO

1892 rows × 5 columns

Cuarto paso del ciclo de vida de datos

A continuacion se va a proceder a realizar el analisis de datos con las funciones SMOTENC y SMOTE, las cuales son:

SMOTE : es tuilizado para un conjunto de datos que contienen caracteristicas numérmicas y categoricas

```
    SMOTENC : a diferencia de SMOTE se envia las variables categorias, es
decir solo las columnas que contienen valores en cadenas, no valores
numericos

estas dos funciones se van a utilizar para un sobremuestreo nominal y
continuo
```

b) Sobremuestre el dataset utilizando SMOTENC y tomando como variable de clase Motivo Desaparción

In [211]:
```
#Imprimimos la columna a utilizar
datos['Motivo Desaparción']
```

Out[211]:
```
0       PROBLEMAS FAMILIARES
1       PROBLEMAS FAMILIARES
2       PROBLEMAS FAMILIARES
3         PROBLEMAS SOCIALES
4       PROBLEMAS FAMILIARES
                ...
1887    PROBLEMAS FAMILIARES
1888    PROBLEMAS FAMILIARES
1889      PROBLEMAS SOCIALES
1890    PROBLEMAS FAMILIARES
1891      PROBLEMAS SOCIALES
Name: Motivo Desaparción, Length: 1892, dtype: object
```

In [212]:
```
#Obtenemos los valores unicos de dicha columna
datos['Motivo Desaparción'].unique()
```

Out[212]:
```
array(['PROBLEMAS FAMILIARES', 'PROBLEMAS SOCIALES',
       'PROBLEMAS PSICOLÓGICOS', 'DISCAPACIDADES Y ENFERMEDADES',
       'EXTRAVIADA', 'PROBLEMA ECÓNOMICOS', 'FALLECIDO',
       'PROBLEMAS ACADÉMICOS', 'PERDIDO', 'FISCALÍA'], dtype=object)
```

In [214]:
```
#Para ver los cambios nuevamente imprimimos la columna
datos['Motivo Desaparción']
#Como se puede observar los cambios han surgido efecto
```

Out[214]:
```
0       PROBLEMAS FAMILIARES
1       PROBLEMAS FAMILIARES
2       PROBLEMAS FAMILIARES
3         PROBLEMAS SOCIALES
4       PROBLEMAS FAMILIARES
                ...
1887    PROBLEMAS FAMILIARES
1888    PROBLEMAS FAMILIARES
1889      PROBLEMAS SOCIALES
1890    PROBLEMAS FAMILIARES
1891      PROBLEMAS SOCIALES
Name: Motivo Desaparción, Length: 1892, dtype: object
```

In [215]:
```
#Obtenemos los valores unicos de dicha columna
datos['Provincia'].unique()
```

Out[215]:
```
array(['PICHINCHA', 'SANTO DOMINGO DE LOS TSACHILAS', 'ESMERALDAS',
       'GUAYAS', 'CHIMBORAZO', 'IMBABURA', 'MANABI', 'AZUAY', 'CARCHI',
       'LOS RIOS', 'TUNGURAHUA', 'SUCUMBIOS', 'EL ORO', 'ORELLANA',
       'NAPO', 'SANTA ELENA', 'LOJA', 'CAÑAR', 'MORONA SANTIAGO',
       'COTOPAXI', 'PASTAZA', 'BOLIVAR', 'ZAMORA CHINCHIPE',
       'ZONA NO DELIMITADA', 'GALAPAGOS'], dtype=object)
```

In [216]:
```
#Obtenemos los valores unicos de dicha columna
datos['Sexo'].unique()
```

Out[216]:
```
array(['HOMBRE', 'MUJER'], dtype=object)
```

In [217]:
```
#Obtenemos los valores unicos de dicha columna
datos['Situación Actual'].unique()
```

Out[217]:
```
array(['DESAPARECIDO', 'ENCONTRADO', 'FALLECIDO'], dtype=object)
```

In [218]:
```
X = datos
```

In [219]:
```
y = datos['Motivo Desaparción']
```

In [220]:
```
y
```

Out[220]:
```
0       PROBLEMAS FAMILIARES
1       PROBLEMAS FAMILIARES
2       PROBLEMAS FAMILIARES
3         PROBLEMAS SOCIALES
4       PROBLEMAS FAMILIARES
                ...
1887    PROBLEMAS FAMILIARES
1888    PROBLEMAS FAMILIARES
1889      PROBLEMAS SOCIALES
1890    PROBLEMAS FAMILIARES
1891      PROBLEMAS SOCIALES
Name: Motivo Desaparción, Length: 1892, dtype: object
```

In [221]:
```
from imblearn.over_sampling import SMOTENC
from collections import Counter
from sklearn.datasets import make_classification
from matplotlib import pyplot
from numpy import where

# summarize class distribution
counter = Counter(y)
print(counter)
# Dentro del parametro categorical_features se le indica las variables que son
#es decir aqui se le va a enviar las columnas que solo contienen letras no las
#aqui se van a indiciar el indice de las variables, la 0,2,3,4
"""
Provincia = 0
Edad Aprox. = 1
Sexo = 2
Motivo Desaparción = 3
Situación Actual = 4
"""
oversample = SMOTENC(categorical_features=[0,2,3,4],k_neighbors=4)
X, y = oversample.fit_resample(X, y)
# summarize the new class distribution
counter = Counter(y)
print(counter)
```

```
Counter({'PROBLEMAS FAMILIARES': 1083, 'PROBLEMAS SOCIALES': 534, 'FALLECID
O': 83, 'DISCAPACIDADES Y ENFERMEDADES': 78, 'PROBLEMAS PSICOLÓGICOS': 38, 'E
XTRAVIADA': 38, 'PROBLEMA ECÓNOMICOS': 13, 'PROBLEMAS ACADÉMICOS': 12, 'FISCA
LÍA': 8, 'PERDIDO': 5})
Counter({'PROBLEMAS FAMILIARES': 1083, 'PROBLEMAS SOCIALES': 1083, 'PROBLEMAS
PSICOLÓGICOS': 1083, 'DISCAPACIDADES Y ENFERMEDADES': 1083, 'EXTRAVIADA': 108
3, 'PROBLEMA ECÓNOMICOS': 1083, 'FALLECIDO': 1083, 'PROBLEMAS ACADÉMICOS': 10
83, 'PERDIDO': 1083, 'FISCALÍA': 1083})
```

In [222]:
```
X
```

Out[222]:

	Provincia	Edad Aprox.	Sexo	Motivo Desaparción	Situación Actual
0	PICHINCHA	17	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
1	PICHINCHA	17	MUJER	PROBLEMAS FAMILIARES	ENCONTRADO
2	SANTO DOMINGO DE LOS TSACHILAS	39	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
3	PICHINCHA	14	MUJER	PROBLEMAS SOCIALES	ENCONTRADO
4	ESMERALDAS	28	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
...	...	...	...	...	...
10825	ESMERALDAS	47	MUJER	PROBLEMAS SOCIALES	ENCONTRADO
10826	PICHINCHA	15	MUJER	PROBLEMAS SOCIALES	ENCONTRADO
10827	NAPO	33	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO
10828	SANTO DOMINGO DE LOS TSACHILAS	27	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO
10829	PICHINCHA	49	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO

10830 rows × 5 columns

In [223]:
```
y
```

Out[223]:
```
0         PROBLEMAS FAMILIARES
1         PROBLEMAS FAMILIARES
2         PROBLEMAS FAMILIARES
3           PROBLEMAS SOCIALES
4         PROBLEMAS FAMILIARES
                  ...
10825       PROBLEMAS SOCIALES
10826       PROBLEMAS SOCIALES
10827       PROBLEMAS SOCIALES
10828       PROBLEMAS SOCIALES
10829       PROBLEMAS SOCIALES
Name: Motivo Desaparción, Length: 10830, dtype: object
```

c) Con el nuevo conjunto de datos realice un nuevo sobremuestreo eligiendo cualquier método (RandomOverSampler,SMOTE,SMOTEN, ADASYN, BorderlineSMOTE, KMeansSMOTE, SVMSMOTE) excepto SMOTENC, y tome como variable de clase a Sexo)

In [224]:
```
#Se realiza una copia de los datos obtenidos con la funcion SMOTENC y se los p
data_aux = X
data_aux
```

Out[224]:

	Provincia	Edad Aprox.	Sexo	Motivo Desaparción	Situación Actual
0	PICHINCHA	17	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
1	PICHINCHA	17	MUJER	PROBLEMAS FAMILIARES	ENCONTRADO
2	SANTO DOMINGO DE LOS TSACHILAS	39	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
3	PICHINCHA	14	MUJER	PROBLEMAS SOCIALES	ENCONTRADO
4	ESMERALDAS	28	HOMBRE	PROBLEMAS FAMILIARES	DESAPARECIDO
...	...	...	...	...	...
10825	ESMERALDAS	47	MUJER	PROBLEMAS SOCIALES	ENCONTRADO
10826	PICHINCHA	15	MUJER	PROBLEMAS SOCIALES	ENCONTRADO
10827	NAPO	33	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO
10828	SANTO DOMINGO DE LOS TSACHILAS	27	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO
10829	PICHINCHA	49	HOMBRE	PROBLEMAS SOCIALES	ENCONTRADO

10830 rows × 5 columns

A contiuacion se va a cambiar todas esas columnas por valores numericos que se especificaran en cada campo

In [225]:
```
#Lo que se va a realizar es cambiar los valores definidos dentro del dataset p
#pueda realizar el ejercicio ya que las cadenas no se puede trabajar, vamos a
#de la siguiente manera entre un rango de 1 a 10
"""
'PROBLEMAS FAMILIARES':1,
'PROBLEMAS SOCIALES':2,
'PROBLEMAS PSICOLÓGICOS':3,
'DISCAPACIDADES Y ENFERMEDADES':4,
'EXTRAVIADA':5,
'PROBLEMA ECÓNOMICOS':6,
'FALLECIDO':7,
'PROBLEMAS ACADÉMICOS':8,
'PERDIDO':9,
'FISCALÍA':10

"""

data_aux['Motivo Desaparción'] = data_aux['Motivo Desaparción'].replace(
    {
        'PROBLEMAS FAMILIARES':1,
        'PROBLEMAS SOCIALES':2,
        'PROBLEMAS PSICOLÓGICOS':3,
        'DISCAPACIDADES Y ENFERMEDADES':4,
        'EXTRAVIADA':5,
        'PROBLEMA ECÓNOMICOS':6,
        'FALLECIDO':7,
        'PROBLEMAS ACADÉMICOS':8,
        'PERDIDO':9,
        'FISCALÍA':10
    }
)
```

In [226]:
```
#Lo que se va a realizar es cambiar los valores definidos dentro del dataset p
#pueda realizar el ejercicio ya que las cadenas no se puede trabajar, vamos a
"""
'PICHINCHA':1,
'SANTO DOMINGO DE LOS TSACHILAS':2,
'ESMERALDAS':3,
'GUAYAS':4,
'CHIMBORAZO':5,
'IMBABURA':6,
'MANABI':7,
'AZUAY':8,
'CARCHI':9,
'LOS RIOS':10,
'TUNGURAHUA':11,
'SUCUMBIOS':12,
'EL ORO':13,
'ORELLANA':14,
'NAPO':15,
'SANTA ELENA':16,
'LOJA':17,
'CAÑAR':18,
'MORONA SANTIAGO':19,
'COTOPAXI':20,
'PASTAZA':21,
'BOLIVAR':22,
'ZAMORA CHINCHIPE':23,
'ZONA NO DELIMITADA':24,
'GALAPAGOS':25

"""

data_aux['Provincia'] = data_aux['Provincia'].replace(
    {
        'PICHINCHA':1,
        'SANTO DOMINGO DE LOS TSACHILAS':2,
        'ESMERALDAS':3,
        'GUAYAS':4,
        'CHIMBORAZO':5,
        'IMBABURA':6,
        'MANABI':7,
        'AZUAY':8,
        'CARCHI':9,
        'LOS RIOS':10,
        'TUNGURAHUA':11,
        'SUCUMBIOS':12,
        'EL ORO':13,
        'ORELLANA':14,
        'NAPO':15,
        'SANTA ELENA':16,
        'LOJA':17,
        'CAÑAR':18,
        'MORONA SANTIAGO':19,
        'COTOPAXI':20,
        'PASTAZA':21,
        'BOLIVAR':22,
        'ZAMORA CHINCHIPE':23,
        'ZONA NO DELIMITADA':24,
```

```
        'GALAPAGOS':25
    }
)
```

In [227]:
```
#Lo que se va a realizar es cambiar los valores definidos dentro del dataset p
#pueda realizar el ejercicio ya que las cadenas no se puede trabajar, vamos a
"""
'HOMBRE':1,
'MUJER':2

"""

data_aux['Sexo'] = data_aux['Sexo'].replace(
    {
        'HOMBRE':1,
        'MUJER':2
    }
)
```

In [228]:
```
#Lo que se va a realizar es cambiar los valores definidos dentro del dataset p
#pueda realizar el ejercicio ya que las cadenas no se puede trabajar, vamos a
"""
'DESAPARECIDO':1,
'ENCONTRADO':2,
'FALLECIDO':3

"""

data_aux['Situación Actual'] = data_aux['Situación Actual'].replace(
    {
        'DESAPARECIDO':1,
        'ENCONTRADO':2,
        'FALLECIDO':3
    }
)
```

In [229]:
```
#Se presenta los datos cambiados a numeros para el uso de la funcion SMOTE
data_aux
```

Out[229]:

	Provincia	Edad Aprox.	Sexo	Motivo Desaparción	Situación Actual
0	1	17	1	1	1
1	1	17	2	1	2
2	2	39	1	1	1
3	1	14	2	2	2
4	3	28	1	1	1
...	...	...	...	...	...
10825	3	47	2	2	2
10826	1	15	2	2	2
10827	15	33	1	2	2
10828	2	27	1	2	2
10829	1	49	1	2	2

10830 rows × 5 columns

In [230]:
```
#A continuacion se va a obtener todas las variables menos las que se va a util
X1= data_aux.drop(['Sexo'],axis=1)
#en la variable y1 se va a obtener los valores de la variable Sexo a utilizar
y1= data_aux['Sexo']
```

In [231]:
```
X1
```

Out[231]:

	Provincia	Edad Aprox.	Motivo Desaparción	Situación Actual
0	1	17	1	1
1	1	17	1	2
2	2	39	1	1
3	1	14	2	2
4	3	28	1	1
...	...	...	...	...
10825	3	47	2	2
10826	1	15	2	2
10827	15	33	2	2
10828	2	27	2	2
10829	1	49	2	2

10830 rows × 4 columns

In [232]:
```
y1
```

Out[232]:
```
0        1
1        2
2        1
3        2
4        1
        ..
10825    2
10826    2
10827    1
10828    1
10829    1
Name: Sexo, Length: 10830, dtype: int64
```

In [240]:
```
X1['Motivo Desaparción'].value_counts()
```

Out[240]:
```
1     1083
2     1083
3     1083
4     1083
5     1083
6     1083
7     1083
8     1083
9     1083
10    1083
Name: Motivo Desaparción, dtype: int64
```

In [ ]:
```

```

In [241]:
```
# Oversample and plot imbalanced dataset with SMOTE
from collections import Counter
from sklearn.datasets import make_classification
from imblearn.over_sampling import SMOTE
from matplotlib import pyplot
from numpy import where

# summarize class distribution
counter = Counter(y1)
print(counter)
# transform the dataset
oversample = SMOTE(k_neighbors=4)
X1, y1 = oversample.fit_resample(X1, y1)
# summarize the new class distribution
counter = Counter(y1)
print(counter)
```

```
Counter({1: 7367, 2: 3463})
Counter({1: 7367, 2: 7367})
```

In [242]:
```
#Se imprime los valores obtenidos en el metodo SMOTE de la variable X1
X1
```

Out[242]:

	Provincia	Edad Aprox.	Motivo Desaparción	Situación Actual
0	1	17	1	1
1	1	17	1	2
2	2	39	1	1
3	1	14	2	2
4	3	28	1	1
...	...	...	...	...
14729	22	11	8	2
14730	1	16	2	2
14731	1	21	10	2
14732	17	41	4	2
14733	4	25	1	1

14734 rows × 4 columns

In [244]:
```
#Se imprime los valores obtenidos en el metodo SMOTE de la variable y1
y1
```

Out[244]:
```
0        1
1        2
2        1
3        2
4        1
        ..
14729    2
14730    2
14731    2
14732    2
14733    2
Name: Sexo, Length: 14734, dtype: int64
```

In [245]:
```
y1.isnull().sum()
```

Out[245]:
```
0
```

In [246]:
```
#Concatenamos los valores que se obtuvieron al aplicar SMOTE
nuevo_dataset = pd.concat([X1, y1],axis=1)
```

In [247]:
```
nuevo_dataset
```

Out[247]:

	Provincia	Edad Aprox.	Motivo Desaparción	Situación Actual	Sexo
0	1	17	1	1	1
1	1	17	1	2	2
2	2	39	1	1	1
3	1	14	2	2	2
4	3	28	1	1	1
...	...	...	...	...	...
14729	22	11	8	2	2
14730	1	16	2	2	2
14731	1	21	10	2	2
14732	17	41	4	2	2
14733	4	25	1	1	2

14734 rows × 5 columns

In [248]:
```
nuevo_dataset['Motivo Desaparción'].value_counts()
```

Out[248]:
```
1     2029
8     2008
3     1706
2     1668
6     1384
4     1273
5     1257
10    1241
7     1085
9     1083
Name: Motivo Desaparción, dtype: int64
```

In [249]:
```
#Verificamos el nuevo dataset que no eixtan valores nulos
nuevo_dataset.isnull().sum()
```

Out[249]:
```
Provincia             0
Edad Aprox.           0
Motivo Desaparción    0
Situación Actual      0
Sexo                  0
dtype: int64
```

Volvemos a los valores originales de las columnas cambiadas

In [250]:

```
"""
Aqui se va volver a los datos originales que tenia el dataset es decir con el

'PROBLEMAS FAMILIARES':1,
'PROBLEMAS SOCIALES':2,
'PROBLEMAS PSICOLÓGICOS':3,
'DISCAPACIDADES Y ENFERMEDADES':4,
'EXTRAVIADA':5,
'PROBLEMA ECÓNOMICOS':6,
'FALLECIDO':7,
'PROBLEMAS ACADÉMICOS':8,
'PERDIDO':9,
'FISCALÍA':10

"""

nuevo_dataset['Motivo Desaparción'] = nuevo_dataset['Motivo Desaparción'].repl
    {
        1:'PROBLEMAS FAMILIARES',
        2: 'PROBLEMAS SOCIALES',
        3: 'PROBLEMAS PSICOLÓGICOS',
        4: 'DISCAPACIDADES Y ENFERMEDADES',
        5: 'EXTRAVIADA',
        6: 'PROBLEMA ECÓNOMICOS',
        7:'FALLECIDO',
        8: 'PROBLEMAS ACADÉMICOS',
        9: 'PERDIDO',
        10: 'FISCALÍA'
    }
)
```

In [251]:

```
"""
'PICHINCHA':1,
'SANTO DOMINGO DE LOS TSACHILAS':2,
'ESMERALDAS':3,
'GUAYAS':4,
'CHIMBORAZO':5,
'IMBABURA':6,
'MANABI':7,
'AZUAY':8,
'CARCHI':9,
'LOS RIOS':10,
'TUNGURAHUA':11,
'SUCUMBIOS':12,
'EL ORO':13,
'ORELLANA':14,
'NAPO':15,
'SANTA ELENA':16,
'LOJA':17,
'CAÑAR':18,
'MORONA SANTIAGO':19,
'COTOPAXI':20,
'PASTAZA':21,
'BOLIVAR':22,
'ZAMORA CHINCHIPE':23,
'ZONA NO DELIMITADA':24,
'GALAPAGOS':25

"""

nuevo_dataset['Provincia'] = nuevo_dataset['Provincia'].replace(
    {
        'PICHINCHA':1,
        'SANTO DOMINGO DE LOS TSACHILAS':2,
        'ESMERALDAS':3,
        'GUAYAS':4,
        'CHIMBORAZO':5,
        'IMBABURA':6,
        'MANABI':7,
        'AZUAY':8,
        'CARCHI':9,
        'LOS RIOS':10,
        'TUNGURAHUA':11,
        'SUCUMBIOS':12,
        'EL ORO':13,
        'ORELLANA':14,
        'NAPO':15,
        'SANTA ELENA':16,
        'LOJA':17,
        'CAÑAR':18,
        'MORONA SANTIAGO':19,
        'COTOPAXI':20,
        'PASTAZA':21,
        'BOLIVAR':22,
        'ZAMORA CHINCHIPE':23,
        'ZONA NO DELIMITADA':24,
        'GALAPAGOS':25
    }
```

```
)
```

In [252]:
```
"""
'HOMBRE':1,
'MUJER':2

"""

nuevo_dataset['Sexo'] = nuevo_dataset['Sexo'].replace(
    {
        1:'HOMBRE',
        2: 'MUJER'
    }
)
```

In [253]:
```
"""
'DESAPARECIDO':1,
'ENCONTRADO':2,
'FALLECIDO':3

"""

nuevo_dataset['Situación Actual'] = nuevo_dataset['Situación Actual'].replace(
    {
        1: 'DESAPARECIDO',
        2: 'ENCONTRADO',
        3: 'FALLECIDO'
    }
)
```

In [254]:
```
nuevo_dataset
```

Out[254]:

	Provincia	Edad Aprox.	Motivo Desaparción	Situación Actual	Sexo
0	1	17	PROBLEMAS FAMILIARES	DESAPARECIDO	HOMBRE
1	1	17	PROBLEMAS FAMILIARES	ENCONTRADO	MUJER
2	2	39	PROBLEMAS FAMILIARES	DESAPARECIDO	HOMBRE
3	1	14	PROBLEMAS SOCIALES	ENCONTRADO	MUJER
4	3	28	PROBLEMAS FAMILIARES	DESAPARECIDO	HOMBRE
...	...	...	...	...	...
14729	22	11	PROBLEMAS ACADÉMICOS	ENCONTRADO	MUJER
14730	1	16	PROBLEMAS SOCIALES	ENCONTRADO	MUJER
14731	1	21	FISCALÍA	ENCONTRADO	MUJER
14732	17	41	DISCAPACIDADES Y ENFERMEDADES	ENCONTRADO	MUJER
14733	4	25	PROBLEMAS FAMILIARES	DESAPARECIDO	MUJER

14734 rows × 5 columns

In [255]:
```
nuevo_dataset.isnull().sum()
```

Out[255]:
```
Provincia             0
Edad Aprox.           0
Motivo Desaparción    0
Situación Actual      0
Sexo                  0
dtype: int64
```

Quinto paso del ciclo de vida de datos

A continuacion se va a proceder a realizar la presentacion de datos obtenido luego de aplicar los métodos pertinentes esto se lo va a hacer presentando la nueva data en un archivo llamado SaucaJosue.csv

In [258]:
```
nuevo_dataset.to_csv('SaucaJosue.csv')
```

In [259]:
```
#Mostrara las 10 filas del nuevo archivo creado
!head -n 50 SaucaJosue.csv
```

```
,Provincia,Edad Aprox.,Motivo DesaparciÃ³n,SituaciÃ³n Actual,Sexo
0,1,17,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
1,1,17,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
2,2,39,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
3,1,14,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
4,3,28,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
5,2,42,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
6,4,16,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
7,4,28,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
8,5,12,PROBLEMAS FAMILIARES,ENCONTRADO,HOMBRE
9,4,2,PROBLEMAS FAMILIARES,DESAPARECIDO,MUJER
10,1,16,PROBLEMAS FAMILIARES,DESAPARECIDO,MUJER
11,6,23,PROBLEMAS SOCIALES,ENCONTRADO,HOMBRE
12,4,21,PROBLEMAS PSICOLÃ“GICOS,ENCONTRADO,MUJER
13,7,17,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
14,7,2,PROBLEMAS FAMILIARES,ENCONTRADO,HOMBRE
15,8,1,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
16,7,15,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
17,9,79,PROBLEMAS FAMILIARES,ENCONTRADO,HOMBRE
18,1,38,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
19,8,23,PROBLEMAS SOCIALES,ENCONTRADO,HOMBRE
20,10,3,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
21,7,37,PROBLEMAS SOCIALES,ENCONTRADO,HOMBRE
22,11,15,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
23,4,16,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
24,12,14,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
25,1,69,DISCAPACIDADES Y ENFERMEDADES,ENCONTRADO,MUJER
26,4,15,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
27,1,16,PROBLEMAS FAMILIARES,ENCONTRADO,HOMBRE
28,4,52,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
29,1,23,DISCAPACIDADES Y ENFERMEDADES,ENCONTRADO,HOMBRE
30,8,64,PROBLEMAS SOCIALES,ENCONTRADO,HOMBRE
31,6,17,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
32,6,29,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
33,6,17,PROBLEMAS SOCIALES,ENCONTRADO,HOMBRE
34,6,45,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
35,4,15,PROBLEMAS FAMILIARES,DESAPARECIDO,MUJER
36,4,13,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
37,1,14,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
38,9,17,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
39,8,20,PROBLEMAS FAMILIARES,DESAPARECIDO,HOMBRE
40,4,15,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
41,8,47,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
42,11,12,EXTRAVIADA,ENCONTRADO,MUJER
43,12,13,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
44,13,22,PROBLEMAS SOCIALES,ENCONTRADO,HOMBRE
45,9,19,PROBLEMAS SOCIALES,ENCONTRADO,MUJER
46,1,2,PROBLEMAS FAMILIARES,ENCONTRADO,MUJER
47,14,12,PROBLEMAS FAMILIARES,ENCONTRADO,HOMBRE
48,15,25,EXTRAVIADA,ENCONTRADO,HOMBRE
```

Sexto paso del ciclo de vida de datos

A continuacion se realiza la toma de desiciones, pero aqui se va a tomar en cuenta una gran pregunta

Responda las siguientes preguntas

Cree que el nuevo dataset creado es un data set confiable para aplicar una mineria de datos y explique el porqué?

```
Considero que la datos obtenidos de la limpieza no son confiables para
aplicar mineria de datos ya que existe un problema de desiequilibramiento de
datos, ya que los algoritmos aplicados a veces puede:

    traer un desbalanceo de las clases
    Falsos valores que viene desde falsos positivos hasta falsos negativos

Esto se entiende como que existen datos desequilbrados, ya que segun el
algoritmo se va a tomar un máximo, incluse al iniicio del ejercicio se
realizo
una media para llenar los datos faltantes, pueda que funcione pero eso seria
controlar la estadisticas y puede aumentar unos valores mas que otros y eso
no esta bien.
A veces se puede descartar informacion util sobre los datos necesarios para
un
muestreo objetivo, esto puede servir para casos de estudio.
```