3. Preprocesamiento en Python

In [14]:

```
import numpy as np
from scipy import stats

import pandas as pd
df = pd.read_csv("caesarian1.txt")

pd.set_option('display.max_rows',None)
#df.head()
#print(df)
df
```

Out[14]:

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
0	22.0	1.0	0.0	Alta	0.0	No
1	26.0	2.0	0.0	Normal	NaN	Si
2	26.0	2.0	1.0	Normal	0.0	No
3	NaN	1.0	0.0	Alta	0.0	No
4	22.0	2.0	0.0	Normal	0.0	Si
5	26.0	1.0	1.0	Baja	0.0	No
6	27.0	2.0	0.0	Normal	0.0	No
7	32.0	NaN	0.0	Normal	0.0	Si
8	28.0	2.0	0.0	Normal	0.0	No
9	27.0	1.0	1.0	Normal	0.0	Si
10	36.0	1.0	0.0	Normal	0.0	No
11	33.0	NaN	NaN	Baja	0.0	Si
12	23.0	1.0	1.0	NaN	0.0	No
13	20.0	1.0	0.0	Normal	1.0	No
14	29.0	1.0	2.0	Baja	1.0	Si
15	25.0	1.0	2.0	Baja	0.0	No
16	25.0	1.0	0.0	Normal	0.0	No
17	20.0	1.0	2.0	Alta	0.0	Si
18	37.0	3.0	0.0	Normal	1.0	Si
19	24.0	1.0	2.0	Baja	1.0	Si
20	26.0	1.0	1.0	Normal	0.0	No
21	33.0	2.0	0.0	Baja	1.0	Si
22	25.0	1.0	1.0	Alta	0.0	No
23	27.0	1.0	0.0	Baja	1.0	Si

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
24	20.0	1.0	0.0	Alta	1.0	Si
25	18.0	1.0	0.0	Normal	0.0	No
26	18.0	1.0	1.0	Alta	1.0	Si
27	30.0	1.0	0.0	Normal	0.0	No
28	32.0	1.0	0.0	Alta	1.0	Si
29	26.0	2.0	1.0	Normal	1.0	No
30	25.0	1.0	0.0	Baja	0.0	No
31	40.0	1.0	0.0	Normal	1.0	Si
32	32.0	2.0	0.0	Alta	1.0	Si
33	27.0	2.0	0.0	Normal	1.0	Si
34	26.0	2.0	2.0	Normal	0.0	Si
35	28.0	3.0	0.0	Alta	0.0	Si
36	33.0	1.0	1.0	Normal	0.0	No
37	31.0	2.0	2.0	Normal	0.0	No
38	31.0	1.0	0.0	Normal	0.0	No
39	26.0	1.0	2.0	Baja	1.0	Si
40	27.0	1.0	0.0	Alta	1.0	Si
41	19.0	1.0	0.0	Normal	0.0	Si
42	36.0	1.0	1.0	Alta	0.0	Si
43	22.0	1.0	0.0	Normal	0.0	Si
44	36.0	4.0	0.0	Alta	1.0	Si
45	28.0	3.0	0.0	Normal	1.0	Si
46	26.0	1.0	0.0	Normal	0.0	No
47	32.0	2.0	0.0	Alta	1.0	Si
48	26.0	2.0	2.0	Normal	0.0	No
49	29.0	2.0	0.0	Baja	1.0	Si
50	33.0	3.0	2.0	Normal	1.0	No
51	21.0	2.0	1.0	Baja	1.0	Si
52	30.0	3.0	2.0	Alta	0.0	No
53	35.0	1.0	1.0	Baja	0.0	No
54	29.0	2.0	0.0	Normal	1.0	Si
55	25.0	2.0	0.0	Normal	0.0	No
56	32.0	3.0	1.0	Baja	1.0	Si

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
57	21.0	1.0	0.0	Baja	0.0	Si
58	26.0	1.0	0.0	Alta	0.0	Si
59	30.0	2.0	1.0	Alta	1.0	Si
60	22.0	1.0	2.0	Alta	0.0	No
61	19.0	1.0	0.0	Normal	0.0	Si
62	32.0	2.0	0.0	Baja	0.0	Si
63	32.0	2.0	0.0	Normal	1.0	Si
64	31.0	1.0	2.0	Alta	1.0	No
65	35.0	2.0	0.0	Normal	0.0	Si
66	28.0	3.0	0.0	Normal	0.0	Si
67	29.0	2.0	0.0	Normal	1.0	No
68	25.0	1.0	0.0	Baja	0.0	Si
69	27.0	2.0	2.0	Baja	0.0	No
70	17.0	1.0	0.0	Baja	0.0	Si
71	29.0	1.0	2.0	Baja	1.0	Si
72	28.0	2.0	0.0	Normal	0.0	No
73	32.0	3.0	0.0	Normal	1.0	No
74	38.0	3.0	2.0	Alta	1.0	Si
75	27.0	2.0	1.0	Normal	0.0	No
76	33.0	4.0	0.0	Normal	0.0	Si
77	29.0	2.0	1.0	Alta	0.0	Si
78	25.0	1.0	2.0	Baja	0.0	Si
79	24.0	2.0	2.0	Normal	0.0	No

IMPUTACION El dataset tiene campos vacios (NAN), como los datos son cualitativos, se procede a calcular la moda para cada columna

In [15]:
```
moda=stats.mode(df)
print(moda)
print(moda[0][0])
m=moda[0][0];
```

```
ModeResult(mode=array([[26.0, 1.0, 0.0, 'Normal', 0.0, 'Si']], dtype=object), count=arra
y([[10, 40, 46, 39, 49, 46]]))
[26.0 1.0 0.0 'Normal' 0.0 'Si']
```

In [16]:
```
df["edad"]=df["edad"].replace(np.nan,m[0])
df["numero_partos"]=df["numero_partos"].replace(np.nan,m[1])
df["tiempo_parto"]=df["tiempo_parto"].replace(np.nan,m[2])
```

```
df["presion_sanguinea"]=df["presion_sanguinea"].replace(np.nan,m[3])
df["problema_corazon"]=df["problema_corazon"].replace(np.nan,m[4])
df
```

Out[16]:

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
0	22.0	1.0	0.0	Alta	0.0	No
1	26.0	2.0	0.0	Normal	0.0	Si
2	26.0	2.0	1.0	Normal	0.0	No
3	26.0	1.0	0.0	Alta	0.0	No
4	22.0	2.0	0.0	Normal	0.0	Si
5	26.0	1.0	1.0	Baja	0.0	No
6	27.0	2.0	0.0	Normal	0.0	No
7	32.0	1.0	0.0	Normal	0.0	Si
8	28.0	2.0	0.0	Normal	0.0	No
9	27.0	1.0	1.0	Normal	0.0	Si
10	36.0	1.0	0.0	Normal	0.0	No
11	33.0	1.0	0.0	Baja	0.0	Si
12	23.0	1.0	1.0	Normal	0.0	No
13	20.0	1.0	0.0	Normal	1.0	No
14	29.0	1.0	2.0	Baja	1.0	Si
15	25.0	1.0	2.0	Baja	0.0	No
16	25.0	1.0	0.0	Normal	0.0	No
17	20.0	1.0	2.0	Alta	0.0	Si
18	37.0	3.0	0.0	Normal	1.0	Si
19	24.0	1.0	2.0	Baja	1.0	Si
20	26.0	1.0	1.0	Normal	0.0	No
21	33.0	2.0	0.0	Baja	1.0	Si
22	25.0	1.0	1.0	Alta	0.0	No
23	27.0	1.0	0.0	Baja	1.0	Si
24	20.0	1.0	0.0	Alta	1.0	Si
25	18.0	1.0	0.0	Normal	0.0	No
26	18.0	1.0	1.0	Alta	1.0	Si
27	30.0	1.0	0.0	Normal	0.0	No
28	32.0	1.0	0.0	Alta	1.0	Si
29	26.0	2.0	1.0	Normal	1.0	No
30	25.0	1.0	0.0	Baja	0.0	No

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
31	40.0	1.0	0.0	Normal	1.0	Si
32	32.0	2.0	0.0	Alta	1.0	Si
33	27.0	2.0	0.0	Normal	1.0	Si
34	26.0	2.0	2.0	Normal	0.0	Si
35	28.0	3.0	0.0	Alta	0.0	Si
36	33.0	1.0	1.0	Normal	0.0	No
37	31.0	2.0	2.0	Normal	0.0	No
38	31.0	1.0	0.0	Normal	0.0	No
39	26.0	1.0	2.0	Baja	1.0	Si
40	27.0	1.0	0.0	Alta	1.0	Si
41	19.0	1.0	0.0	Normal	0.0	Si
42	36.0	1.0	1.0	Alta	0.0	Si
43	22.0	1.0	0.0	Normal	0.0	Si
44	36.0	4.0	0.0	Alta	1.0	Si
45	28.0	3.0	0.0	Normal	1.0	Si
46	26.0	1.0	0.0	Normal	0.0	No
47	32.0	2.0	0.0	Alta	1.0	Si
48	26.0	2.0	2.0	Normal	0.0	No
49	29.0	2.0	0.0	Baja	1.0	Si
50	33.0	3.0	2.0	Normal	1.0	No
51	21.0	2.0	1.0	Baja	1.0	Si
52	30.0	3.0	2.0	Alta	0.0	No
53	35.0	1.0	1.0	Baja	0.0	No
54	29.0	2.0	0.0	Normal	1.0	Si
55	25.0	2.0	0.0	Normal	0.0	No
56	32.0	3.0	1.0	Baja	1.0	Si
57	21.0	1.0	0.0	Baja	0.0	Si
58	26.0	1.0	0.0	Alta	0.0	Si
59	30.0	2.0	1.0	Alta	1.0	Si
60	22.0	1.0	2.0	Alta	0.0	No
61	19.0	1.0	0.0	Normal	0.0	Si
62	32.0	2.0	0.0	Baja	0.0	Si
63	32.0	2.0	0.0	Normal	1.0	Si

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
64	31.0	1.0	2.0	Alta	1.0	No
65	35.0	2.0	0.0	Normal	0.0	Si
66	28.0	3.0	0.0	Normal	0.0	Si
67	29.0	2.0	0.0	Normal	1.0	No
68	25.0	1.0	0.0	Baja	0.0	Si
69	27.0	2.0	2.0	Baja	0.0	No
70	17.0	1.0	0.0	Baja	0.0	Si
71	29.0	1.0	2.0	Baja	1.0	Si
72	28.0	2.0	0.0	Normal	0.0	No
73	32.0	3.0	0.0	Normal	1.0	No
74	38.0	3.0	2.0	Alta	1.0	Si
75	27.0	2.0	1.0	Normal	0.0	No
76	33.0	4.0	0.0	Normal	0.0	Si
77	29.0	2.0	1.0	Alta	0.0	Si
78	25.0	1.0	2.0	Baja	0.0	Si
79	24.0	2.0	2.0	Normal	0.0	No

DISCRETIZAR, Si 1, no 0, valores cuantitativos

In [17]:
```
df["class"]=df["class"].replace("Si",1)
df["class"]=df["class"].replace("No",0)

df
```

Out[17]:

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
0	22.0	1.0	0.0	Alta	0.0	0
1	26.0	2.0	0.0	Normal	0.0	1
2	26.0	2.0	1.0	Normal	0.0	0
3	26.0	1.0	0.0	Alta	0.0	0
4	22.0	2.0	0.0	Normal	0.0	1
5	26.0	1.0	1.0	Baja	0.0	0
6	27.0	2.0	0.0	Normal	0.0	0
7	32.0	1.0	0.0	Normal	0.0	1
8	28.0	2.0	0.0	Normal	0.0	0
9	27.0	1.0	1.0	Normal	0.0	1
10	36.0	1.0	0.0	Normal	0.0	0

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
11	33.0	1.0	0.0	Baja	0.0	1
12	23.0	1.0	1.0	Normal	0.0	0
13	20.0	1.0	0.0	Normal	1.0	0
14	29.0	1.0	2.0	Baja	1.0	1
15	25.0	1.0	2.0	Baja	0.0	0
16	25.0	1.0	0.0	Normal	0.0	0
17	20.0	1.0	2.0	Alta	0.0	1
18	37.0	3.0	0.0	Normal	1.0	1
19	24.0	1.0	2.0	Baja	1.0	1
20	26.0	1.0	1.0	Normal	0.0	0
21	33.0	2.0	0.0	Baja	1.0	1
22	25.0	1.0	1.0	Alta	0.0	0
23	27.0	1.0	0.0	Baja	1.0	1
24	20.0	1.0	0.0	Alta	1.0	1
25	18.0	1.0	0.0	Normal	0.0	0
26	18.0	1.0	1.0	Alta	1.0	1
27	30.0	1.0	0.0	Normal	0.0	0
28	32.0	1.0	0.0	Alta	1.0	1
29	26.0	2.0	1.0	Normal	1.0	0
30	25.0	1.0	0.0	Baja	0.0	0
31	40.0	1.0	0.0	Normal	1.0	1
32	32.0	2.0	0.0	Alta	1.0	1
33	27.0	2.0	0.0	Normal	1.0	1
34	26.0	2.0	2.0	Normal	0.0	1
35	28.0	3.0	0.0	Alta	0.0	1
36	33.0	1.0	1.0	Normal	0.0	0
37	31.0	2.0	2.0	Normal	0.0	0
38	31.0	1.0	0.0	Normal	0.0	0
39	26.0	1.0	2.0	Baja	1.0	1
40	27.0	1.0	0.0	Alta	1.0	1
41	19.0	1.0	0.0	Normal	0.0	1
42	36.0	1.0	1.0	Alta	0.0	1
43	22.0	1.0	0.0	Normal	0.0	1

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
44	36.0	4.0	0.0	Alta	1.0	1
45	28.0	3.0	0.0	Normal	1.0	1
46	26.0	1.0	0.0	Normal	0.0	0
47	32.0	2.0	0.0	Alta	1.0	1
48	26.0	2.0	2.0	Normal	0.0	0
49	29.0	2.0	0.0	Baja	1.0	1
50	33.0	3.0	2.0	Normal	1.0	0
51	21.0	2.0	1.0	Baja	1.0	1
52	30.0	3.0	2.0	Alta	0.0	0
53	35.0	1.0	1.0	Baja	0.0	0
54	29.0	2.0	0.0	Normal	1.0	1
55	25.0	2.0	0.0	Normal	0.0	0
56	32.0	3.0	1.0	Baja	1.0	1
57	21.0	1.0	0.0	Baja	0.0	1
58	26.0	1.0	0.0	Alta	0.0	1
59	30.0	2.0	1.0	Alta	1.0	1
60	22.0	1.0	2.0	Alta	0.0	0
61	19.0	1.0	0.0	Normal	0.0	1
62	32.0	2.0	0.0	Baja	0.0	1
63	32.0	2.0	0.0	Normal	1.0	1
64	31.0	1.0	2.0	Alta	1.0	0
65	35.0	2.0	0.0	Normal	0.0	1
66	28.0	3.0	0.0	Normal	0.0	1
67	29.0	2.0	0.0	Normal	1.0	0
68	25.0	1.0	0.0	Baja	0.0	1
69	27.0	2.0	2.0	Baja	0.0	0
70	17.0	1.0	0.0	Baja	0.0	1
71	29.0	1.0	2.0	Baja	1.0	1
72	28.0	2.0	0.0	Normal	0.0	0
73	32.0	3.0	0.0	Normal	1.0	0
74	38.0	3.0	2.0	Alta	1.0	1
75	27.0	2.0	1.0	Normal	0.0	0
76	33.0	4.0	0.0	Normal	0.0	1

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
77	29.0	2.0	1.0	Alta	0.0	1
78	25.0	1.0	2.0	Baja	0.0	1
79	24.0	2.0	2.0	Normal	0.0	0

CATEGORIZAR, Rural=R, Urbana=U, Valores Cualitativos

In [18]:
```
df["presion_sanguinea"]=df["presion_sanguinea"].replace("Alta","A")
df["presion_sanguinea"]=df["presion_sanguinea"].replace("Normal","N")
df["presion_sanguinea"]=df["presion_sanguinea"].replace("Baja","B")

df
```

Out[18]:

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
0	22.0	1.0	0.0	A	0.0	0
1	26.0	2.0	0.0	N	0.0	1
2	26.0	2.0	1.0	N	0.0	0
3	26.0	1.0	0.0	A	0.0	0
4	22.0	2.0	0.0	N	0.0	1
5	26.0	1.0	1.0	B	0.0	0
6	27.0	2.0	0.0	N	0.0	0
7	32.0	1.0	0.0	N	0.0	1
8	28.0	2.0	0.0	N	0.0	0
9	27.0	1.0	1.0	N	0.0	1
10	36.0	1.0	0.0	N	0.0	0
11	33.0	1.0	0.0	B	0.0	1
12	23.0	1.0	1.0	N	0.0	0
13	20.0	1.0	0.0	N	1.0	0
14	29.0	1.0	2.0	B	1.0	1
15	25.0	1.0	2.0	B	0.0	0
16	25.0	1.0	0.0	N	0.0	0
17	20.0	1.0	2.0	A	0.0	1
18	37.0	3.0	0.0	N	1.0	1
19	24.0	1.0	2.0	B	1.0	1
20	26.0	1.0	1.0	N	0.0	0
21	33.0	2.0	0.0	B	1.0	1
22	25.0	1.0	1.0	A	0.0	0
23	27.0	1.0	0.0	B	1.0	1

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
24	20.0	1.0	0.0	A	1.0	1
25	18.0	1.0	0.0	N	0.0	0
26	18.0	1.0	1.0	A	1.0	1
27	30.0	1.0	0.0	N	0.0	0
28	32.0	1.0	0.0	A	1.0	1
29	26.0	2.0	1.0	N	1.0	0
30	25.0	1.0	0.0	B	0.0	0
31	40.0	1.0	0.0	N	1.0	1
32	32.0	2.0	0.0	A	1.0	1
33	27.0	2.0	0.0	N	1.0	1
34	26.0	2.0	2.0	N	0.0	1
35	28.0	3.0	0.0	A	0.0	1
36	33.0	1.0	1.0	N	0.0	0
37	31.0	2.0	2.0	N	0.0	0
38	31.0	1.0	0.0	N	0.0	0
39	26.0	1.0	2.0	B	1.0	1
40	27.0	1.0	0.0	A	1.0	1
41	19.0	1.0	0.0	N	0.0	1
42	36.0	1.0	1.0	A	0.0	1
43	22.0	1.0	0.0	N	0.0	1
44	36.0	4.0	0.0	A	1.0	1
45	28.0	3.0	0.0	N	1.0	1
46	26.0	1.0	0.0	N	0.0	0
47	32.0	2.0	0.0	A	1.0	1
48	26.0	2.0	2.0	N	0.0	0
49	29.0	2.0	0.0	B	1.0	1
50	33.0	3.0	2.0	N	1.0	0
51	21.0	2.0	1.0	B	1.0	1
52	30.0	3.0	2.0	A	0.0	0
53	35.0	1.0	1.0	B	0.0	0
54	29.0	2.0	0.0	N	1.0	1
55	25.0	2.0	0.0	N	0.0	0
56	32.0	3.0	1.0	B	1.0	1

	edad	numero_partos	tiempo_parto	presion_sanguinea	problema_corazon	class
57	21.0	1.0	0.0	B	0.0	1
58	26.0	1.0	0.0	A	0.0	1
59	30.0	2.0	1.0	A	1.0	1
60	22.0	1.0	2.0	A	0.0	0
61	19.0	1.0	0.0	N	0.0	1
62	32.0	2.0	0.0	B	0.0	1
63	32.0	2.0	0.0	N	1.0	1
64	31.0	1.0	2.0	A	1.0	0
65	35.0	2.0	0.0	N	0.0	1
66	28.0	3.0	0.0	N	0.0	1
67	29.0	2.0	0.0	N	1.0	0
68	25.0	1.0	0.0	B	0.0	1
69	27.0	2.0	2.0	B	0.0	0
70	17.0	1.0	0.0	B	0.0	1
71	29.0	1.0	2.0	B	1.0	1
72	28.0	2.0	0.0	N	0.0	0
73	32.0	3.0	0.0	N	1.0	0
74	38.0	3.0	2.0	A	1.0	1
75	27.0	2.0	1.0	N	0.0	0
76	33.0	4.0	0.0	N	0.0	1
77	29.0	2.0	1.0	A	0.0	1
78	25.0	1.0	2.0	B	0.0	1
79	24.0	2.0	2.0	N	0.0	0

In [ ]: