Creado por:

Isabel Maniega

In [1]:
```
# pip install pandas
```

In [2]:
```
# pip install scikit-learn
```

In [3]:
```
# pip install seaborn
```

In [5]:
```
import pandas as pd
from sklearn.datasets import fetch_openml
import seaborn as sns
from sklearn.impute import SimpleImputer
import numpy as np
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
```

Cargar el titanic con sklearn

In [8]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[8]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	

In [ ]:
```
# df.to_csv("Titanic_all.csv")
```

Mostrar las columnas sin datos

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 13 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   pclass     1309 non-null   float64
 1   name       1309 non-null   object
 2   sex        1309 non-null   category
 3   age        1046 non-null   float64
 4   sibsp      1309 non-null   float64
 5   parch      1309 non-null   float64
 6   ticket     1309 non-null   object
 7   fare       1308 non-null   float64
 8   cabin      295 non-null    object
 9   embarked   1307 non-null   category
 10  boat       486 non-null    object
 11  body       121 non-null    float64
 12  home.dest  745 non-null    object
dtypes: category(2), float64(6), object(5)
memory usage: 115.4+ KB
```

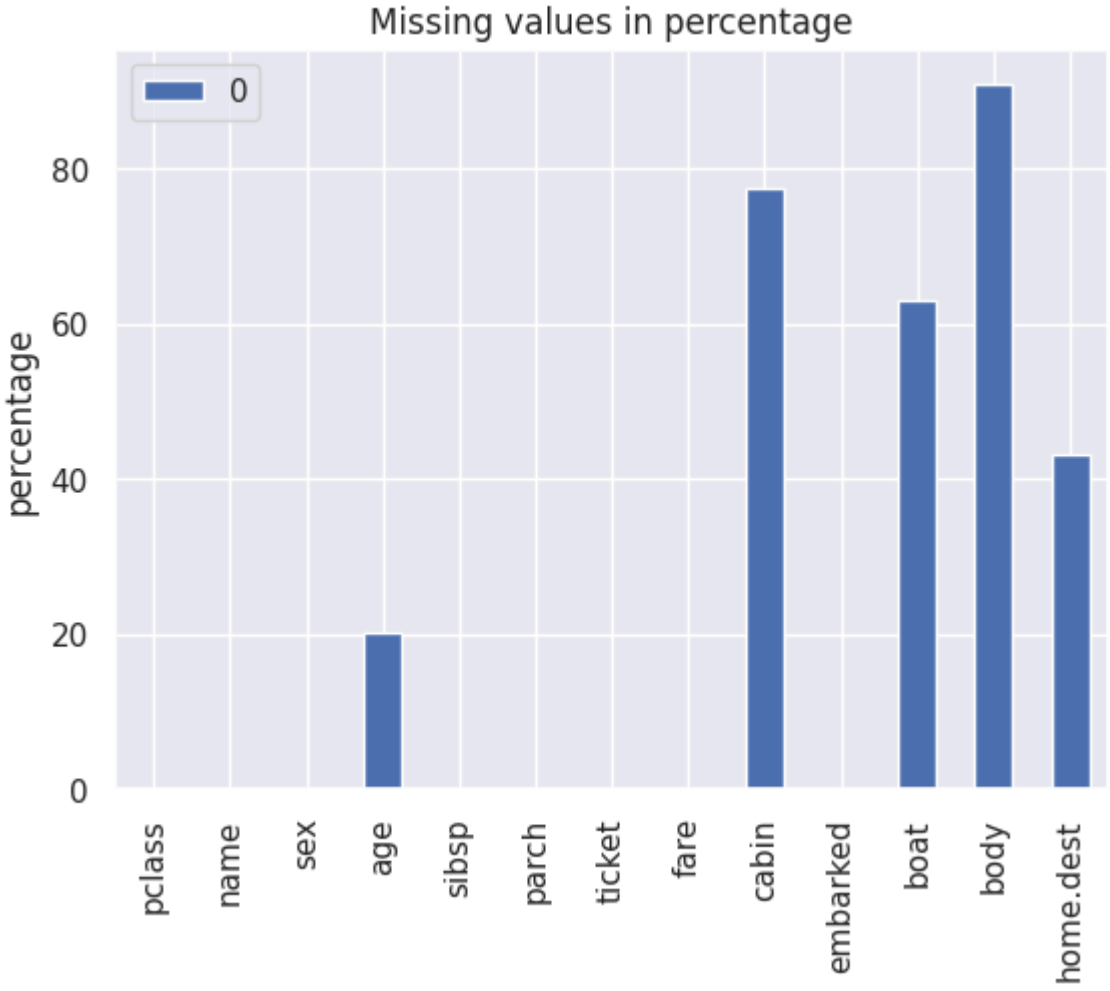
In [10]:
```
df.isnull().sum()
```

Out[10]:
```
pclass          0
name            0
sex             0
age           263
sibsp           0
parch           0
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest     564
dtype: int64
```

In [11]:
```
# Visusalización de los datos

sns.set()
miss_vals = pd.DataFrame(df.isnull().sum()/ len(df)*100)
miss_vals.plot(kind="bar",
               title="Missing values in percentage",
               ylabel="percentage")
```

Out[11]:
```
<Axes: title={'center': 'Missing values in percentage'}, ylabel='percent
age'>
```



Procedimiento para valores nulos

Existen dos maneras:

- Eliminar la columna
- Asignamos a los valores la media, mediana, moda, etc

Eliminación

- Eliminamos los valores nulos:

In [12]:
```
print(f"Size of the dataset: {df.shape}")
df.drop(["cabin", "boat", "body", "home.dest"], axis=1, inplace=True)
df.dropna(inplace=True)
print(f"Size of the dataset: {df.shape}")
```

```
Size of the dataset: (1309, 13)
Size of the dataset: (1043, 9)
```

Sustitución

- Sustituir por el valor más común (media):

In [13]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[13]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	

In [14]:
```
df.describe()
```

Out[14]:

	pclass	age	sibsp	parch	fare	body
count	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000	121.000000
mean	2.294882	29.881135	0.498854	0.385027	33.295479	160.809917
std	0.837836	14.413500	1.041658	0.865560	51.758668	97.696922
min	1.000000	0.166700	0.000000	0.000000	0.000000	1.000000
25%	2.000000	21.000000	0.000000	0.000000	7.895800	72.000000
50%	3.000000	28.000000	0.000000	0.000000	14.454200	155.000000
75%	3.000000	39.000000	1.000000	0.000000	31.275000	256.000000
max	3.000000	80.000000	8.000000	9.000000	512.329200	328.000000

In [15]:
```
print(f"Número de valores nulos de la columna edad: {df.age.isnull().sum(
```

```
Número de valores nulos de la columna edad: 263
```

In [17]:
```
df["age"].fillna(df["age"].mean())
print(f"Número de valores nulos de la columna edad: {df.age.isnull().sum(
```

```
Número de valores nulos de la columna edad: 0
```

- Otra opción: Simple transformación con Sklearn

In [18]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[18]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	

In [19]:
```
print(f"Número de valores nulos de la columna edad: {df.age.isnull().sum(
```

```
Número de valores nulos de la columna edad: 263
```

In [20]:
```
imp = SimpleImputer(strategy="mean")
df["age"] = imp.fit_transform(df[["age"]])
print(f"Número de valores nulos de la columna edad: {df.age.isnull().sum(
```

```
Número de valores nulos de la columna edad: 0
```

In [21]:
```
print("Tipos de datos con valores Nulos:")
for col in df.columns[df.isnull().any()]:
    print(col, df[col][df[col].isnull()].values[0])
```

```
Tipos de datos con valores Nulos:
fare nan
cabin None
embarked nan
boat None
body nan
home.dest None
```

- Modificamos los None:

In [22]:
```
imp = SimpleImputer(missing_values=None, strategy="most_frequent")
df["cabin"] = imp.fit_transform(df[["cabin"]])
print(f"Número de valores nulos de la columna cabina: {df.cabin.isnull().
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[22], line 2
      1 imp = SimpleImputer(missing_values=None, strategy="most_frequent")
----> 2 df["cabin"] = imp.fit_transform(df[["cabin"]])
      3 print(f"Número de valores nulos de la columna cabina: {df.cabin.isnull().sum()}")

File ~/.local/lib/python3.10/site-packages/pandas/core/frame.py:4299, in DataFrame.__setitem__(self, key, value)
   4296     self._setitem_array([key], value)
   4297 else:
   4298     # set column
-> 4299     self._set_item(key, value)

File ~/.local/lib/python3.10/site-packages/pandas/core/frame.py:4512, in DataFrame._set_item(self, key, value)
   4502 def _set_item(self, key, value) -> None:
   4503     """
   4504     Add series to DataFrame in specified column.
   4505
   (...)
   4510     ensure homogeneity.
   4511     """
-> 4512     value, refs = self._sanitize_column(value)
   4514     if (
   4515         key in self.columns
   4516         and value.ndim == 1
   4517         and not isinstance(value.dtype, ExtensionDtype)
   4518     ):
   4519         # broadcast across multiple columns if necessary
   4520         if not self.columns.is_unique or isinstance(self.columns, MultiIndex):

File ~/.local/lib/python3.10/site-packages/pandas/core/frame.py:5254, in DataFrame._sanitize_column(self, value)
   5252 if is_list_like(value):
   5253     com.require_length_match(value, self.index)
-> 5254 arr = sanitize_array(value, self.index, copy=True, allow_2d=True)
   5255 if (
   5256     isinstance(value, Index)
   5257     and value.dtype == "object"
   (...)
   5260     # TODO: Remove kludge in sanitize_array for string mode when enforcing
   5261     # this deprecation
   5262     warnings.warn(
   5263         "Setting an Index with object dtype into a DataFrame will stop "
   5264         "inferring another dtype in a future version. Cast the Index "
   (...)
   5267         stacklevel=find_stack_level(),
   5268     )

File ~/.local/lib/python3.10/site-packages/pandas/core/construction.py:606, in sanitize_array(data, index, dtype, copy, allow_2d)
    604 subarr = data
```

```
    605 if data.dtype == object:
--> 606     subarr = maybe_infer_to_datetimelike(data)
    607     if (
    608         object_index
    609         and using_pyarrow_string_dtype()
    610         and is_string_dtype(subarr)
    611     ):
    612         # Avoid inference when string option is set
    613         subarr = data

File ~/.local/lib/python3.10/site-packages/pandas/core/dtypes/cast.py:118
2, in maybe_infer_to_datetimelike(value)
   1179     raise TypeError(type(value))  # pragma: no cover
   1180 if value.ndim != 1:
   1181     # Caller is responsible
-> 1182     raise ValueError(value.ndim)  # pragma: no cover
   1184 if not len(value):
   1185     return value

ValueError: 2
```

In [23]:
```
def get_parameters(df):
    parameters = {}
    for col in df.columns[df.isnull().any()]:
        if df[col].dtype == "float64" or df[col].dtype == "int64" or df[c
            strategy = "mean"
        else:
            strategy = "most_frequent"
        missing_values = df[col][df[col].isnull()].values[0]
        parameters[col] = {"missing_values": missing_values, "strategy":
    return parameters
get_parameters(df)
```

Out[23]:
```
{'fare': {'missing_values': nan, 'strategy': 'mean'},
 'cabin': {'missing_values': None, 'strategy': 'most_frequent'},
 'embarked': {'missing_values': nan, 'strategy': 'most_frequent'},
 'boat': {'missing_values': None, 'strategy': 'most_frequent'},
 'body': {'missing_values': nan, 'strategy': 'mean'},
 'home.dest': {'missing_values': None, 'strategy': 'most_frequent'}}
```

In [24]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[24]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	

In [25]:

```
parameters = get_parameters(df)

for col, param in parameters.items():
    missing_values = param["missing_values"]
    strategy = param["strategy"]
    imp = SimpleImputer(missing_values=missing_values, strategy=strategy)
    df[col] = imp.fit_transform(df[[col]])

df.isnull().sum()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[25], line 7
      5     strategy = param["strategy"]
      6     imp = SimpleImputer(missing_values=missing_values, strategy=strategy)
----> 7     df[col] = imp.fit_transform(df[[col]])
      9 df.isnull().sum()

File ~/.local/lib/python3.10/site-packages/pandas/core/frame.py:4299, in DataFrame.__setitem__(self, key, value)
   4296     self._setitem_array([key], value)
   4297 else:
   4298     # set column
-> 4299     self._set_item(key, value)

File ~/.local/lib/python3.10/site-packages/pandas/core/frame.py:4512, in DataFrame._set_item(self, key, value)
   4502 def _set_item(self, key, value) -> None:
   4503     """
   4504     Add series to DataFrame in specified column.
   4505
   (...)
   4510     ensure homogeneity.
   4511     """
-> 4512     value, refs = self._sanitize_column(value)
   4514     if (
   4515         key in self.columns
   4516         and value.ndim == 1
   4517         and not isinstance(value.dtype, ExtensionDtype)
   4518     ):
   4519         # broadcast across multiple columns if necessary
   4520         if not self.columns.is_unique or isinstance(self.columns, MultiIndex):

File ~/.local/lib/python3.10/site-packages/pandas/core/frame.py:5254, in DataFrame._sanitize_column(self, value)
   5252 if is_list_like(value):
   5253     com.require_length_match(value, self.index)
-> 5254 arr = sanitize_array(value, self.index, copy=True, allow_2d=True)
   5255 if (
   5256     isinstance(value, Index)
   5257     and value.dtype == "object"
   (...)
   5260     # TODO: Remove kludge in sanitize_array for string mode when enforcing
   5261     # this deprecation
   5262     warnings.warn(
   5263         "Setting an Index with object dtype into a DataFrame will stop "
   5264         "inferring another dtype in a future version. Cast the Index "
   (...)
   5267         stacklevel=find_stack_level(),
   5268     )

File ~/.local/lib/python3.10/site-packages/pandas/core/construction.py:606, in sanitize_array(data, index, dtype, copy, allow_2d)
```

```
    604 subarr = data
    605 if data.dtype == object:
--> 606     subarr = maybe_infer_to_datetimelike(data)
    607     if (
    608         object_index
    609         and using_pyarrow_string_dtype()
    610         and is_string_dtype(subarr)
    611     ):
    612         # Avoid inference when string option is set
    613         subarr = data

File ~/.local/lib/python3.10/site-packages/pandas/core/dtypes/cast.py:118
2, in maybe_infer_to_datetimelike(value)
   1179     raise TypeError(type(value))  # pragma: no cover
   1180 if value.ndim != 1:
   1181     # Caller is responsible
-> 1182     raise ValueError(value.ndim)  # pragma: no cover
   1184 if not len(value):
   1185     return value

ValueError: 2
```

In [26]:
```
df.head()
```

Out[26]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	

Crear nuevas características (Feature Engineering)

- Sibsp: pasajeros que viajan con hermanos
- Parch: viajeros que viajan con niños

Calculamos el número de pasajeros que viajan solos:

In [27]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[27]:

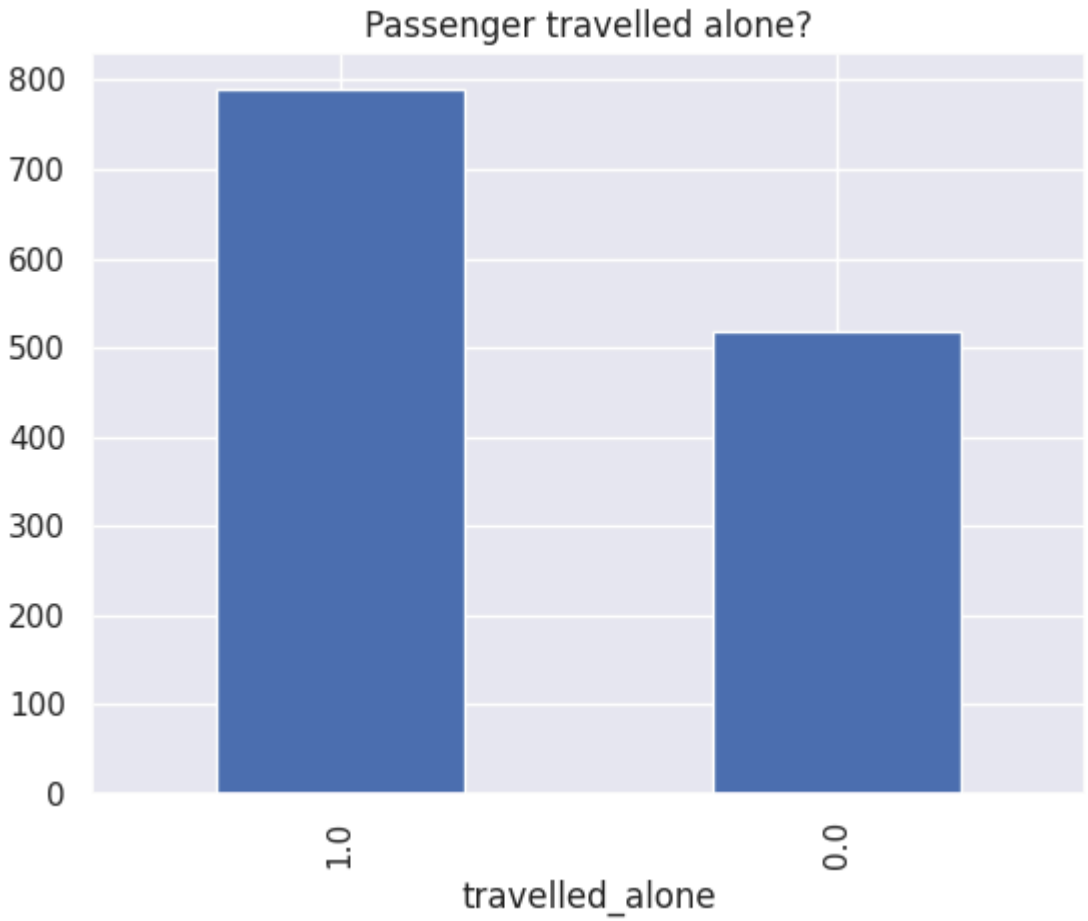
	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	

In [28]:
```
df["family"] = df["sibsp"] + df["parch"]

df.loc[df["family"] > 0, "travelled_alone"] = 0
df.loc[df["family"] == 0, "travelled_alone"] = 1

df["travelled_alone"].value_counts().plot(title="Passenger travelled alon
```

Out[28]:
```
<Axes: title={'center': 'Passenger travelled alone?'}, xlabel='travelled
_alone'>
```



Encode categorical features

- scikit-learn: OneHotEncoder()
- pandas: get_dummies()

In [29]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[29]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	
1	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	
2	1.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	
3	1.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	

In [30]:
```
df[["female", "male"]] = OneHotEncoder().fit_transform(df[['sex']]).toarr
df[["sex", "female", "male"]]
```

Out[30]:

	sex	female	male
0	female	1.0	0.0
1	male	0.0	1.0
2	female	1.0	0.0
3	male	0.0	1.0
4	female	1.0	0.0
...	...	...	...
1304	female	1.0	0.0
1305	female	1.0	0.0
1306	male	0.0	1.0
1307	male	0.0	1.0
1308	male	0.0	1.0

1309 rows × 3 columns

Eliminaremos uno de las columnas para evitar la colinealidad

In [31]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]
df["sex"] = OneHotEncoder().fit_transform(df[['sex']]).toarray()[:, 1]
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_openml.py:968: FutureWarning: The default value of `parser` will change from `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silence this warning. Therefore, an `ImportError` will be raised from 1.4 if the dataset is dense and pandas is not installed. Note that the pandas parser may return different data types. See the Notes Section in fetch_openml's API doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation with empty or all-NA entries is deprecated. In a future version, this will no longer exclude empty or all-NA columns when determining the result dtypes. To retain the old behavior, exclude the relevant entries before the concat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[31]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	1.0	Allen, Miss. Elisabeth Walton	0.0	29.0000	0.0	0.0	24160	211.3375	B5	S
1	1.0	Allison, Master. Hudson Trevor	1.0	0.9167	1.0	2.0	113781	151.5500	C22 C26	S
2	1.0	Allison, Miss. Helen Loraine	0.0	2.0000	1.0	2.0	113781	151.5500	C22 C26	S
3	1.0	Allison, Mr. Hudson Joshua Creighton	1.0	30.0000	1.0	2.0	113781	151.5500	C22 C26	S
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	0.0	25.0000	1.0	2.0	113781	151.5500	C22 C26	S

0 == female; 1== male

- Pandas:

In [32]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]

df["sex"] = pd.get_dummies(df["sex"], drop_first=True, dtype=float)
df.head()
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_openml.py:968: FutureWarning: The default value of `parser` will change from `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silence this warning. Therefore, an `ImportError` will be raised from 1.4 if the dataset is dense and pandas is not installed. Note that the pandas parser may return different data types. See the Notes Section in fetch_openml's API doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation with empty or all-NA entries is deprecated. In a future version, this will no longer exclude empty or all-NA columns when determining the result dtypes. To retain the old behavior, exclude the relevant entries before the concat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

Out[32]:

	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	1.0	Allen, Miss. Elisabeth Walton	0.0	29.0000	0.0	0.0	24160	211.3375	B5	S
1	1.0	Allison, Master. Hudson Trevor	1.0	0.9167	1.0	2.0	113781	151.5500	C22 C26	S
2	1.0	Allison, Miss. Helen Loraine	0.0	2.0000	1.0	2.0	113781	151.5500	C22 C26	S
3	1.0	Allison, Mr. Hudson Joshua Creighton	1.0	30.0000	1.0	2.0	113781	151.5500	C22 C26	S
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	0.0	25.0000	1.0	2.0	113781	151.5500	C22 C26	S

0 == female; 1== male

Encoding all categorical features

In [33]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]

cat_cols = df.select_dtypes(include=["category"]).columns
print(f"Columnas Categoricas: {cat_cols}")
```

```
Columnas Categoricas: Index(['sex', 'embarked'], dtype='object')
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

In [34]:
```
for col in cat_cols:
    fill_value = df[col].mode()[0]
    df[col].fillna(fill_value, inplace=True)

    append_to = list(df[col].unique())

    print(append_to)

    df[append_to] = OneHotEncoder().fit_transform(df[[col]]).toarray()

    df.drop(col, axis=1, inplace=True)
    df.drop(append_to[0], axis=1, inplace=True)

print(df.columns)
df[["male", "C", "Q"]].head()
```

```
['female', 'male']
['S', 'C', 'Q']
Index(['pclass', 'name', 'age', 'sibsp', 'parch', 'ticket', 'fare', 'cabi
n',
       'boat', 'body', 'home.dest', 'male', 'C', 'Q'],
      dtype='object')
```

```
/tmp/ipykernel_139809/1246407664.py:3: FutureWarning: A value is trying to
be set on a copy of a DataFrame or Series through chained assignment using
an inplace method.
The behavior will change in pandas 3.0. This inplace method will never wor
k because the intermediate object on which we are setting values always be
haves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  df[col].fillna(fill_value, inplace=True)
/tmp/ipykernel_139809/1246407664.py:3: FutureWarning: A value is trying to
be set on a copy of a DataFrame or Series through chained assignment using
an inplace method.
The behavior will change in pandas 3.0. This inplace method will never wor
k because the intermediate object on which we are setting values always be
haves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  df[col].fillna(fill_value, inplace=True)
```

Out[34]:

	male	C	Q
0	0.0	0.0	1.0
1	1.0	0.0	1.0
2	0.0	0.0	1.0
3	1.0	0.0	1.0
4	0.0	0.0	1.0

In [35]:
```
df.head()
```

Out[35]:

	pclass	name	age	sibsp	parch	ticket	fare	cabin	boat	body	h
0	1.0	Allen, Miss. Elisabeth Walton	29.0000	0.0	0.0	24160	211.3375	B5	2	NaN	
1	1.0	Allison, Master. Hudson Trevor	0.9167	1.0	2.0	113781	151.5500	C22 C26	11	NaN	Cl
2	1.0	Allison, Miss. Helen Loraine	2.0000	1.0	2.0	113781	151.5500	C22 C26	None	NaN	Cl
3	1.0	Allison, Mr. Hudson Joshua Creighton	30.0000	1.0	2.0	113781	151.5500	C22 C26	None	135.0	Cl
4	1.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	25.0000	1.0	2.0	113781	151.5500	C22 C26	None	NaN	Cl

MinMaxScaler

MinMaxScaler() pone todos los valores númericos de 0 a 1:

In [36]:
```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]

num_cols = df.select_dtypes(include=["int64", "int32", "float64"]).column
print(num_cols)
```

```
Index(['pclass', 'age', 'sibsp', 'parch', 'fare', 'body'], dtype='object')
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

In [37]:
```
for col in num_cols:
    fill_value = df[col].mean()
    df[col].fillna(fill_value, inplace=True)
```

```
minmax = MinMaxScaler()

df[num_cols] = minmax.fit_transform(df[num_cols])
df[num_cols]
```

```
/tmp/ipykernel_139809/2406708159.py:3: FutureWarning: A value is trying to
be set on a copy of a DataFrame or Series through chained assignment using
an inplace method.
The behavior will change in pandas 3.0. This inplace method will never wor
k because the intermediate object on which we are setting values always be
haves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  df[col].fillna(fill_value, inplace=True)
```

Out[37]:

	pclass	age	sibsp	parch	fare	body
0	0.0	0.361169	0.000	0.000000	0.412503	0.488715
1	0.0	0.009395	0.125	0.222222	0.295806	0.488715
2	0.0	0.022964	0.125	0.222222	0.295806	0.488715
3	0.0	0.373695	0.125	0.222222	0.295806	0.409786
4	0.0	0.311064	0.125	0.222222	0.295806	0.488715
...	...	...	...	...	...	...
1304	1.0	0.179540	0.125	0.000000	0.028213	1.000000
1305	1.0	0.372206	0.125	0.000000	0.028213	0.488715
1306	1.0	0.329854	0.000	0.000000	0.014102	0.926606
1307	1.0	0.336117	0.000	0.000000	0.014102	0.488715
1308	1.0	0.361169	0.000	0.000000	0.015371	0.488715

1309 rows × 6 columns

StandardScaler

StandardScaler() poner todos los valores tengan una media de 0 y de desviación de 1

In [38]:

```
df = fetch_openml("titanic", version=1, as_frame=True)["data"]

num_cols = df.select_dtypes(include=["int64", "int32", "float64"]).column
print(num_cols)
```

```
Index(['pclass', 'age', 'sibsp', 'parch', 'fare', 'body'], dtype='object')
```

```
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
openml.py:968: FutureWarning: The default value of `parser` will change fr
om `'liac-arff'` to `'auto'` in 1.4. You can set `parser='auto'` to silenc
e this warning. Therefore, an `ImportError` will be raised from 1.4 if the
dataset is dense and pandas is not installed. Note that the pandas parser
may return different data types. See the Notes Section in fetch_openml's A
PI doc for details.
  warn(
/home/isabelmaniega/.local/lib/python3.10/site-packages/sklearn/datasets/_
arff_parser.py:200: FutureWarning: The behavior of DataFrame concatenation
with empty or all-NA entries is deprecated. In a future version, this will
no longer exclude empty or all-NA columns when determining the result dtyp
es. To retain the old behavior, exclude the relevant entries before the co
ncat operation.
  frame = pd.concat(dfs, ignore_index=True)
```

In [39]:
```
for col in num_cols:
    fill_value = df[col].mean()
    df[col].fillna(fill_value, inplace=True)

ss = StandardScaler()

df[num_cols] = ss.fit_transform(df[num_cols])
df[num_cols].head()
```

```
/tmp/ipykernel_139809/2660934333.py:3: FutureWarning: A value is trying to
be set on a copy of a DataFrame or Series through chained assignment using
an inplace method.
The behavior will change in pandas 3.0. This inplace method will never wor
k because the intermediate object on which we are setting values always be
haves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  df[col].fillna(fill_value, inplace=True)
```

Out[39]:

	pclass	age	sibsp	parch	fare	body
0	-1.546098	-0.068420	-0.479087	-0.445000	3.442480	0.000000
1	-1.546098	-2.249092	0.481288	1.866526	2.286476	0.000000
2	-1.546098	-2.164974	0.481288	1.866526	2.286476	0.000000
3	-1.546098	0.009230	0.481288	1.866526	2.286476	-0.872539
4	-1.546098	-0.379021	0.481288	1.866526	2.286476	0.000000

In [40]:
```
df[num_cols].describe()
```

Out[40]:

	pclass	age	sibsp	parch	fare	
count	1.309000e+03	1.309000e+03	1.309000e+03	1.309000e+03	1.309000e+03	1.30
mean	-1.737003e-16	1.519878e-16	-8.142201e-18	1.628440e-17	4.342507e-17	-2.57
std	1.000382e+00	1.000382e+00	1.000382e+00	1.000382e+00	1.000382e+00	1.00
min	-1.546098e+00	-2.307330e+00	-4.790868e-01	-4.449995e-01	-6.437751e-01	-5.40
25%	-3.520907e-01	-6.119712e-01	-4.790868e-01	-4.449995e-01	-4.911082e-01	0.00
50%	8.419164e-01	2.758687e-16	-4.790868e-01	-4.449995e-01	-3.643001e-01	0.00
75%	8.419164e-01	3.974806e-01	4.812878e-01	-4.449995e-01	-3.906640e-02	0.00
max	8.419164e-01	3.891737e+00	7.203909e+00	9.956864e+00	9.262219e+00	5.65

Usando el método de describe() podemos ver la media y la desvicación estandar de las columnas escaladas.

La media no parece ser igual a 0 pero, de hecho 4.342507e-17 es igual 0,00000000000000043425. Esto es tan cercano a 0 que puede considerarse igual a 0. Lo mismo ocurre con la desviación estándar que es tan cercana a 1 que puede considerarse igual a 1.

Creado por:

Isabel Maniega