Primera práctica: Adquisición y adecuación de base de datos.

Objetivo

Aplicar el proceso de adecuación de la *data* (limpieza) a los conjuntos de datos proporcionados, para utilizarlos en los modelos de *machine learning*. Los datos se proporcionan en formato csv y se descargan en los siguientes enlaces:

- https://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data
- https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

Pasos a seguir (orientativo): Prepara un conjunto de datos confiable.

- Realizar la carga de datos desde Python (Google Colab) o R-studio.
- Análisis descriptivo de los datos que incluya las medidas de tendencia central.
- Tratamiento de *missing*.
- Identificación y eliminación de errores.
- Datos duplicados.
- Comentar los resultados obtenidos.
- Otros comentarios que parezcan adecuados.

El enlace: https://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data contiene la información necesaria sobre los datos, la variable respuesta es “Class name”. Los datos tratan información relacionada con el voto en las elecciones americanas.

Cargue data:

```
nRowsRead = None
COLUMNS2 = ['Class Name:','handicapped-infants', 'water-project-cost-sharing',
            'adoption-of-the-budget-resolution', 'education_num', 'marital',
            'physician-fee-freeze ', 'el-salvador-aid', ' religious-groups-in-schools',
            'anti-satellite-test-ban', 'aid-to-nicaraguan-contras', 'mx-missile',
            'immigration', ' synfuels-corporation-cutback', 'education-spending',
            'superfund-right-to-sue', 'crime', 'duty-free-exports', 'export-administration-act-south-africa']
PATH = "https://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data"
df2 = pd.read_csv(PATH, delimiter=',', nrows = nRowsRead, names=COLUMNS2)
df2.dataframeName = 'house-votes-84.data.csv'
nRow, nCol = df2.shape
print(f'Hay {nRow} filas y {nCol} columnas')
```

```
Hay 435 filas y 19 columnas
```

```
[51]  df2.head(5)
```

	Class Name:	handicapped-infants	water-project-cost-sharing	adoption-of-the-budget-resolution	education_num	marital	physician-fee-freeze	el-salvador-aid	religious-groups-in-schools	anti-satellite-test-ban	aid-to-nicaraguan-contras
0	republican	n	y	n	y	y	y	n	n	n	y
1	republican	n	y	n	y	y	y	n	n	n	n
2	democrat	?	y	y	?	y	y	n	n	n	n
3	democrat	n	y	y	n	?	y	n	n	n	n
4	democrat	y	y	y	n	y	y	n	n	n	n

El enlace contiene los datos e incluye ocho variables categóricas y seis variables continuas: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

Cargue data:

```
nRowsRead = None
COLUMNS = ['age','workclass', 'fnlwgt', 'education', 'education_num', 'marital',
           'occupation', 'relationship', 'race', 'sex', 'capital_gain', 'capital_loss',
           'hours_week', 'native_country', 'label']
# tic-tac-toe.data.csv may have more rows in reality, but we are only loading/previewing the first 1000 rows
#PATH = "https://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.data"
PATH = "https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data"
df1 = pd.read_csv(PATH, delimiter=',', nrows = nRowsRead, names=COLUMNS)
df1.dataframeName = 'house-votes-84.data.csv'
nRow, nCol = df1.shape
print(f'Hay {nRow} filas y {nCol} columnas')
```

```
Hay 32561 filas y 15 columnas
```

```
[50]  df1.head(5)
```

	age	workclass	fnlwgt	education	education_num	marital	occupation	relationship	race	sex	capital_gain	capital_loss
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0

Comentarios sobre la evaluación

- Máximo de cinco páginas.
- Se pueden usar R o Python.
- Se deben comentar los resultados obtenidos y el código.

Deadline

Según la guía docente/organización.