Programación para Data Science

Unidad 6: Preprocesamiento de datos en Python - Ejercicios

Cargad los datos del fichero `bank_edited.csv` en un dataframe. Este conjunto de datos recoge datos sobre una campaña de márketing de un banco portugués. El conjunto original puede encontrarse en el repositorio de datos de Machine Learning de la UC Irvine (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing), pero el conjunto que usaremos tiene algunas modificaciones.

Nota: revisad la documentación de la función `read_csv` (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html) para ver de qué parámetros disponemos para ajustar el proceso de carga de datos

Ejercicio 1

Los valores del estado civil (atributo `marital`) contienen errores tipográficos e incluyen el uso de distintas nomenclaturas. En este ejercicio, unificaremos la nomenclatura de los valores de este atributo. **(1.5 puntos)**

a) ¿Cuántos valores distintos toma el atributo `marital` en el conjunto de datos? Mostrad cuáles son estos valores.

In [ ]:
```
# Respuesta
```

b) Unificad los valores del atributo `marital` para sean únicamente: "single", "married" o "divorced".

In [ ]:
```
# Respuesta
```

c) Comprobad que la unificación que habéis realizado es correcta.

In [ ]:
```
# Respuesta
```

Ejercicio 2

(2 puntos)

a) ¿Qué columnas contienen valores perdidos?

In [ ]:
```
# Respuesta
```

b) ¿Cuántas muestras tienen al menos un valor perdido en cualquiera de sus atributos?

In [ ]:
```
# Respuesta
```

c) ¿Cuántas muestras tienen al menos tres valores perdidos?

In [ ]:
```
# Respuesta
```

d) Eliminad las muestras que tengan algún valor perdido en cualquiera de los atributos

In [ ]:
```
# Respuesta
```

Ejercicio 3

Cread dos nuevas columnas en el dataframe con el resultado de discretizar el atributo `duration`: **(1.5 puntos)**

a) La columna `disc_1` contendrá la discretización en 10 intervalos que contengan el mismo número de muestras cada uno.

In [ ]:
```
# Respuesta
```

b) La columna `disc_2` contendrá la discretización en 10 intervalos del mismo tamaño.

In [ ]:
```
# Respuesta
```

Ejercicio 4

Crea 4 nuevos atributos binarios (`balance_Q1`, `balance_Q2`, `balance_Q3`, `balance_Q4`) que indiquen en qué cuartil se encuentra el valor de balance (`balance`) de cada muestra. Así, para una muestra con un valor de balance en el segundo cuartil, el atributo `balance_Q2` será `1` y los atributos `balance_Q1`, `balance_Q3` y `balance_Q4` serán `0`. **(2 puntos)**

In [ ]:
```
# Respuesta
```

Ejercicio 5

Selecciona aleatoriamente un subconjunto de 20 muestras del dataframe, de manera que el dataframe resultante esté balanceado con respeto al atributo `default`. Es decir, el nuevo dataframe deber tener el mismo número de muestras para cada posible valor del atributo `default`. **(1.5 puntos)**

In [ ]:
```
# Respuesta
```

Ejercicio 6

El atributo `poutcome` contiene información sobre si el cliente del banco contrató un depósito a largo plazo con el banco. Si disponéis únicamente de los datos del dataframe y tuvierais que escoger una única columna para precedir si el cliente contrató el depósito, ¿qué columna escogeríais?**(1.5 puntos)**

Nota: obviamente, no podéis escoger la columna poutcome

Pista: Quizás la función corr (http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html) de pandas puede ser de utilidad.

In [ ]:
```
# Respuesta
```