Programación para *Data Science*

Unidad 6: Preprocesamiento de datos en Python - Ejercicios y preguntas

Ejercicio 0

Cargad los datos del fichero `winequality-white_edited.csv` en un *dataframe*. Este conjunto de datos recoge datos sobre muestras de vino del norte de Portugal. El conjunto original puede encontrarse en el repositorio de datos de Machine Learning de la UC Irvine (http://archive.ics.uci.edu/ml/datasets/Wine+Quality), pero el conjunto que usaremos tiene algunas modificaciones.

Mostrad el número de muestras del conjunto de datos y los nombres de los atributos (columnas). **(0.5 puntos)**

```
In [1]: # Respuesta
```

Ejercicio 1

El valor de la densidad del vino (atributo `density`) es un valor cercano a 1. **(1 punto)**

a) ¿Cuántas muestras contienen valores de densidad erróneos?

```
In [2]: # Respuesta
```

b) Sustituid todos los valores de densidad incorrectos, asumiendo que representan la parte decimal de un valor con parte entera 0. Por ejemplo, un valor 558 tendría que transformarse en 0.558.

Pista: notad que todos los valores erróneos tienen tres cifras.

```
In [3]: # Respuesta
```

Ejercicio 2

(1.5 puntos)

a) ¿Qué columnas contienen valores perdidos?

```
In [4]: # Respuesta
```

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

```
In [5]: # Respuesta
```

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

```
In [6]: # Respuesta
```

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

```
In [7]: # Respuesta
```

Ejercicio 3

Una de las columnas del conjunto de datos contiene información totalmente redundante. Describid cómo detectar qué columna es y realizad el proceso de detección. Eliminad la columna del conjunto de datos. **(1.5 puntos)**

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
```

Pregunta 1

En el Notebook de ejemplo de este módulo hemos visto cómo escalar los datos utilizando el `StandardScaler` del módulo `preprocessing` de sklearn. Describid alguna otra alternativa para escalar los datos y las propiedades que esta alternativa aporta. **(1 punto)**

Respuesta:

Ejercicio 4

Escalad los valores de los atributos `residual sugar` y `free sulfur dioxide` siguiendo la estrategia descrita en la pregunta anterior. **(1 punto)**

In [8]:
```
# Respuesta
```

Ejercicio 5

Cread 4 nuevos atributos binarios (`sulphates_Q1`, `sulphates_Q2`, `sulphates_Q3`, `sulphates_Q4`) que indiquen en qué cuartil se encuentra el valor de sulfatos (`sulphates`) de cada muestra. Así, para una muestra con un valor de sulfatos en el segundo cuartil, el atributo `sulphates_Q2` será `1` y los atributos `sulphates_Q1`, `sulphates_Q3` y `sulphates_Q4` serán 0. **(2 puntos)**

In [9]:
```
# Respuesta
```

Ejercicio 6

Selecciona aleatoriamente un subconjunto de cien muestras y crea un *dataframe* con estas muestras, incluyendo únicamente los valores de los atributos `alcohol`, `quality` y `chlorides`. **(1.5 puntos)**

In [10]:
```
# Respuesta
```