# prog_datasci_6_preproc_entrega

October 19, 2020

# 1 Programación para *Data Science*

## 1.1 Unidad 6: Preprocesamiento de datos en Python

# 2 Introducción

En este Notebook encontraréis el conjunto de **actividades evaluables** como PEC de la asignatura.

Or recordamos que os miréis antes de abordar este notebook los ejercicios para practicar. Éstos no puntúan para la PEC, pero os recomendamos que los intentéis resolver como parte del proceso de aprendizaje. Encontraréis ejemplos de posibles soluciones a los ejercicios al propio notebook, pero es importante que intentéis resolverlos vosotros antes de consultar las soluciones. Las soluciones os permitirán validar vuestras respuestas, así como ver alternativas de resolución de las actividades. También os animamos a preguntar cualquier duda que surja sobre la resolución de los **ejercicios para practicar** en el foro del aula.

En relación a las actividades evaluables de este notebook, veréis que cada una de ellas tiene asociada una puntuación que indica el peso que tiene la actividad sobre la nota de la PEC. Adicionalmente, hay un ejercicio opcional, que no tiene puntuación dentro de la PEC, pero que se valora al final del semestre de cara a conceder las matrículas de honor y redondear las notas finales. Podéis sacar la máxima nota de la PEC sin necesidad de hacer este ejercicio! El objetivo de este ejercicio es que sirva como pequeño reto para los estudiantes que quieran profundizar en el contenido de la asignatura.

Además, veréis que todas las actividades tienen una etiqueta que indica los recursos necesarios para llevarla a cabo. Hay tres posibles etiquetas:

- NM **Sólo materiales**: las herramientas necesarias para realizar la actividad se pueden encontrar en los materiales de la asignatura.
- EG **Consulta externa guiada**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, pero el enunciado contiene indicaciones de dónde o cómo encontrar la información adicional necesaria para resolver la actividad.
- EI **Consulta externa independente**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, y el enunciado puede no incluir la descripción de dónde o cómo encontrar esta información adicional. Será necesario que el estudiante busque esta información utilizando los recursos que se han explicado en la asignatura.

Es importante notar que estas etiquetas no indican el nivel de dificultad del ejercicio, sino únicamente la necesidad de consulta de documentación externa para su resolución. Además, recordad

que las **etiquetas son informativas**, pero podréis consultar referencias externas en cualquier momento (aunque no se indique explícitamente) o puede ser que podáis hacer una actividad sin consultar ningún tipo de documentación. Por ejemplo, para resolver una actividad que sólo requiera los materiales de la asignatura, puedéis consultar referencias externas si queréis, ya sea tanto para ayudaros en la resolución como para ampliar el conocimiento!

En cuanto a la consulta de documentación externa en la resolución de los ejercicios, recordad **citar siempre la bibliografía utilizada** para resolver cada actividad.

# 3 Ejercicios para la PEC

A continuación, los **ejercicios y preguntas teóricas que debe completar en esta PEC** y que forman parte de la evaluación de esta unidad.

# 4 Ejercicio 1

El patrón de consumo de alcohol es un rasgo característico de cada país y muchas veces está asociado a ciertos aspectos sociales. En el siguiente archivo `drinks.csv`, encontraréis la cantidad de consumo de alcohol por país considerando tres tipos de bebidas: cerveza, vino y licor y, además, el total de litros de alcohol puro consumido. El conjunto original puede encontrarse en el repositorio de datos FiveThirtyEigth, pero el conjunto de datos que utilizaremos tiene algunas modificaciones.

**Nota**: revisad la documentación de la función read_csv para ver que parámetros disponéis para ajustar el proceso de carga de los datos. Una vez importados los datos, realizad las siguientes operaciones:

a) Importad el fichero `drinks.csv`de la carpeta de datos en un dataframe. Mostrad por pantalla el nombre total de medidas y el nombre de las variables. Mostrad las 10 primeras entradas.

**(0,5 puntos)** NM

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

b) Observad la codificación de la variable `continent`. ¿Cuántos valores diferentes toma el atributo `continent` en el conjunto de datos? Unificad los valores de la variable `continent` para obtener los siguientes valores: “Asia”, “Africa”, “Oceania”, “Europa” y “America”.

**(1 punto)** NM

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

c) Busca los valores anómalos o sin sentido en las variables asociadas al consumo de alcohol (cerveza, licor, vino y total). Pensad que valores pueden ser anómalos o sin sentido en el conjunto de datos proporcionado. Tened en cuenta, entre otras cosas, que el consumo de cerveza, vino y licor respecto el total para buscar posibles valores anómalos. Si no hay consumo de ninguna de las tres bebidas alcohólicas, el total de alcohol consumido tiene que ser cero. A partir de este criterio, substituye todos los valores anómalos por `NaN`, excepto aquellos valores anómalos que se ajustan al criterio anterior. En estos casos, estos valores anómalos tienen que ser substituidos por cero.

**(1.5 puntos)** EI

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

d) Mostrad por pantalla todos los valores perdidos. Imputad los valores perdidos de cada variable por la mediana por continente.

**Nota**: Consultad primero los tipos de variables de cada una de las columnas y valorad si alguna de las variables del dataframe tiene que transformarse.

**(1.5 puntos)** EG

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

e) Comprueba si hay valores perdidos en el dataframe.

**(0,5 puntos)** NM

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

### 4.1 Ejercicio 2

En este ejercicio trabajaremos con el conjunto de datos derivado del ejercicio anterior.

a) Calculad los siguientes valores estadísticos: media aritmética, máximo, mínimo, cuartiles (25%,50% y 75%) y desviación típica de todas las variables del dataframe por continente.

**(1 punto)** NM

[26]:
```
# Respuesta
```

b) Para comparar las diferentes medidas de una variable tipificaremos los datos para reducir el impacto de los valores atípicos o marginales. Este método transforma una variable según una determinada distribución. Utilizad sklearn.preprocessing para tipificar nuestro conjunto de datos mediante una transformación no lineal basada en cuartiles y siguiendo una distribución uniforme.

**Nota**: Consultad el siguiente artículo para ver la importancia de la estandarización de los datos en el rendimiento de los algoritmos de aprendizaje automatizado.

**(1,5 puntos)** EG

[28]:
```
# Respuesta
```

c) La OMS (Organización Mundial de la Salud) quiere determinar si la bebida que más contribuye en el total de alcohol puro consumido es la misma para todos los continentes. Para realizarlo, calcula la correlación de la variable `total_litres_of_pure_alcohol` con el resto de variables mediante el coeficiente de correlación de Pearson. Utilizando el conjunto de datos del apartado anterior, determina que bebida contribuye más en el total de alcohol puro consumido por continente. Razonad la respuesta.

**(1,5 puntos)** EG

[30]:
```
# Respuesta
```

d) El atributo `continent` del conjunto de datos no está balanceado, no tiene el mismo número de observaciones para cada una de las clases. Los algoritmos de clasificación tienden a favorecer las clases mayoritarias obteniendo métricas de clasificación sesgadas por clase. Determina que clase es la minoritaria y cuál debería de ser el número de observaciones de cada clase para tener el conjunto de datos balanceado. Selecciona aleatoriamente un subconjunto mínimo de muestras del dataframe, de forma que el dataframe resultante esté balanceado respecto a la variable `continent`.

**(1 punto)** NM

[32]:
```
# Respuesta
```

### 4.1.1 Ejercicio Opcional

Los valores atípicos son valores que se alejan del resto de variables y, en la mayoría de los casos, son valores que pueden perturbar el resultado final. Por esta razón es necesario identificarlos y, generalmente, eliminarlos del análisis. Utilizad la función `IsolationForest` de la librería sklearn.ensemble para determinar los valores atípicos considerando una proporción de valores atípicos en el conjunto de datos de 0,2. Determina cuantos valores atípicos hay en el conjunto de datos derivados del ejercicio 1 y que países deberían de ser eliminados del conjunto de datos final.

EI

[35]:
```
# Respuesta
```