Programació per a *Data Science*

Unitat 2: Introducció a l'anàlisi de dades en Python

En aquest Notebook trobareu el conjunt d'activitats avaluables com PAC de l'assignatura. Veureu que cadascuna d'elles té associada una puntuació, que indica el pes que té l'activitat sobre la nota final de la PAC. Addicionalment, hi ha un exercici opcional, que no té puntuació dins de la PAC, però que es valora a la fi de l'semestre de cara a concedir les matrícules d'honor i arrodonir les notes finals. Podreu treure la màxima nota de la PAC sense necessitat de fer aquest exercici. L'objectiu d'aquest exercici és que serveixi com a petit repte per als estudiants que vulguin aprofundir en el contingut de l'assignatura.

Veureu que totes les activitats de la PAC tenen una etiqueta, que indica els recursos necessaris per tal de dur-la a terme. Hi ha tres possibles etiquetes:

- `NM` **Només materials**: les eines necessàries per a realitzar l'activitat es poden trobar als materials de l'assignatura.
- `EG` **Consulta externa guiada**: l'activitat pot requerir fer ús d'eines que no es troben als materials de l'assignatura, però l'enunciat conté indicacions d'on o com trobar la informació addicional necessària per resoldre l'activitat.
- `EI` **Consulta externa independent**: l'activitat pot requerir fer ús d'eines que no es troben als materials de l'assignatura, i l'enunciat pot no incloure la descripció d'on o com trobar aquesta informació addicional. Caldrà que l'estudiant busqui aquesta informació fent servir els recursos que s'han explicat a l'assignatura.

És important notar que aquestes etiquetes no indiquen el nivell de dificultat de l'exercici, sinó únicament la necessitat de consulta de documentació externa per a la seva resolució. A més, recordeu que les **etiquetes són informatives**, però podeu consultar referències externes sempre que vulgueu (encara que no s'indiqui explícitament) o pot ser que pugueu fer una activitat sense consultar cap mena de documentació. Per exemple, per resoldre una activitat que només requereixi els materials de l'assignatura, podeu consultar referències externes si voleu, ja sigui tant per ajudar-vos en la reva resolució com per ampliar el coneixement!

Pel que fa a la consulta de documentació externa en la resolució dels exercicis, recordeu **citar sempre la bibliografia utilitzada** per a resoldre cada activitat.

Exercici 1

Carrega el conjunt de dades wine incorporat a la llibreria `sklearn`. Per resoldre aquest exercici, només podeu utilitzar la secció `data`.

En aquest conjunt de dades, hi ha vins de diferents categories, però de moment no sabem a quina categoria perteneix cadascun.

Donat aquest conjunt de dades, en un restaurant ens demanen endevinar en quants grups podríem categoritzar els vins per així emmagatzemar-los en cellers diferents, intentant que això requereixi el mínim nombre de sales però que al mateix temps permeti representar el millor possible les diferències entre aquests vins.

1) L'algoritme K-Means ens permet agrupar els vins en diferents clústers. No obstant això, caldrà avaluar el seu rendiment per separar les diferents categories en relació al nombre de clústers utilitzats. Així, podem representar el nombre d'habitacions simplement ajustant el nombre de clústers en l'algoritme K-means. Quin és el menor nombre d'habitacions que permeten emmagatzemar els vins amb la major diferència entre els vins de cada habitació? Implementa la solució i mostra el resultat gràficament. **(2 punts)**

Nota: En els exercicis resolts hem vist com fer servir la funció `fit` de K-means per separar les dades. Tingueu en compte que, un cop fet el *fit*, la funció `inertia_` ens permet quantificar el rendiment de l'algorisme per separar els grups (separarabilitat).

2) Seria possible utilitzar una tècnica d'anàlisi de dades **supervisada** per resoldre aquest problema? Per què? **(1 punt)** `NM`

In [10]:

```
# Resposta
```

Resposta

Exercici 2

A partir d'aquestes dades crea un *dataframe*, incloent una columna amb el nom *class*, que indiqui el tipus de vi, és a dir la variable `target`. La resta de columnes correspondran amb les característiques o `feature_names`. Finalment, mostra per pantalla la

capçalera del *dataframe* i els descriptius estadístics del *dataset* o conjunt de dades. **(1 punt)** NM

In [11]:

```
### Resposta
```

Exercici 3

Ara que teniu el conjunt de dades organitzades, traceu la distribució de cadascuna de les característiques per a cada classe de vi individualment. Per a això podeu utilitzar historgrames com hem vist al Notebook de teoria o podeu triar utilitzar un [diagrama de violí de la llibreria Seaborn] (https://seaborn.pydata.org/generated/seaborn.violinplot.html#seaborn.violinplot). Assegureu-vos de mostrar una sola figura amb múltiples panells, dedicant un panell a cada característica dels vins (ex. Magnesi). A cada panell es mostraran les distribucions de valors per a cada classe de vi. **(1 punt)** NM

In [12]:

```
# Resposta
```

Exercici 4

L'**anàlisi discriminant lineal (LDA per les sigles en anglès)** és un mètode de classificació supervisat de variables qualitatives en què dos o més grups són coneguts a priori i noves observacions es classifiquen en un d'ells en funció de les seves característiques. Fent ús del teorema de Bayes, LDA estima la probabilitat que una observació, donat un determinat valor dels predictors (caractereistiques), pertanyi a cadascuna de les classes de la variable qualitativa: $P(I = k|X = x)$. Finalment s'assigna l'observació a la classe k per la qual la probabilitat predita és més gran.

En aquest exercici, haureu de crear un model basat en LDA per a classificar les classes de vi segons les característiques del conjunt de dades.

3.1 En primer lloc, haureu d'estandarditzar les dades utilitzant el mètode StandardScaler de sklearn.

3.2- A continuació, haureu de dividir les dades en conjunts d'entrenament (*training*) i avaluació (*test*), amb una mida de la mostra per a test del 33%. 3.3- A continuació, entreneu el model LDA amb la porció de dades d'entrenament i calculeu la precisió de el model. Tindreu que fer servir el classificador LDA de sklearn per entrenar el model i el mètode score () per avaluar-lo. Escriviu una breu explicació que indiqui el que significa obtenir tal precisió.

(3 punts) EG

In [13]:

```
# Resposta

import numpy as np
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

In [14]:

```
# Elimineu la columna de classe del conjunt de dades, ja que seran les nostres dades de
destinació:
X = wine_data_df.drop(['class'], axis=1)
y = wine_data_df['class']
```

Exercici 5

Ara que teniu el model entrenat, haureu d'aplicar la transformació del LDA al conjunt de dades de manera que us permeti reduir la seva dimensionalitat a 2 components. Un cop fet això, haureu de representar la transformació resultant gràficament utilitzant un diagrama de dispersió. Per a això, haureu generar dues figures, una amb les dades transformades del conjunt d'entrenament (*training*) i una altra del conjunt d'avaluació (*test*).

Nota: No oblideu donar un color a cada punt de dades segons la classe de vi a la qual correspon.

(2 punts) EI

In [15]:

In [15]:

```
# Resposta
```

Exercici opcional

Ara que hem après algunes coses sobre els tipus de vi i les seves característiques, imagina que un amic està prenent una copa de vi i et llegeix el següent etiquetatge:

```
alcohol =                              14.23
malic_acid =                            1.71
ash =                                   2.43
alcalinity_of_ash =                    15.60
magnesium =                           127.00
total_phenols =                         2.80
flavanoids =                            3.06
nonflavanoid_phenols =                  0.28
proanthocyanins =                       2.29
color_intensity =                       5.64
hue =                                   1.04
od280_od315_of_diluted_wines =          3.92
proline =                            1065.00
```

Podries endevinar quina classe de vi està prenent el teu amic? Quina és la probabilitat d'equivocar-te?

Nota: Utilitza el model LDA entrenat per endevinar la classe de vi.

EG

In [17]:

```
# Resposta
```