Nous allons étendre la notion de voisins à des objets à plusieurs caractéristiques.

L'idée est d'utiliser un grand nombre de données afin "d'apprendre à la machine" à résoudre un certain type de problème . Cette idée d'apprentissage automatique ne date pas d'hier, puisque le terme de ***machine learning*** a été utilisé pour la première fois par l'informaticien américain Arthur Samuel en ***1959***.
Pourquoi le machine learning est tant "à la mode" depuis quelques années ? Simplement parce que le nerf de la guerre dans les algorithmes de machine learning est la qualité et la quantité des données (les données qui permettront à la machine d'apprendre à résoudre un problème), or, avec le développement d'internet, il est relativement simple de trouver des données sur n'importe quel sujet (on parle de "big data").
À noter aussi l'importance des stratégies mises en place par les GAFAM (Google, Apple, Facebook, Amazon et Microsoft) afin de récupérer un grand nombre de données concernant leurs clients. Ces données sont très souvent utilisées pour "nourrir" des algorithmes de machine learning (comment, d'après vous, Amazon arrive à proposer à ces clients des "suggestions d'achats" souvent très pertinentes ?)

Afin de travailler sur un exemple, nous allons utiliser un jeu de données portant sur des fleurs, connu dans le monde du machine learning : le jeu de données "iris".
En 1936, Edgar Anderson a collecté des données sur 3 espèces d'iris : "iris setosa", "iris virginica" et "iris versicolor" .

iris setosa

pétales

Pour chaque iris étudié, il a mesuré (en cm) :

- la largeur des sépales
- la longueur des sépales
- la largeur des pétales
- la longueur des pétales
- l'espèce ("iris setosa", "iris virginica" ou "iris versicolor")

Par souci de simplification, nous nous intéresserons uniquement à la largeur et à la longueur des pétales.

sépales

iris virginica

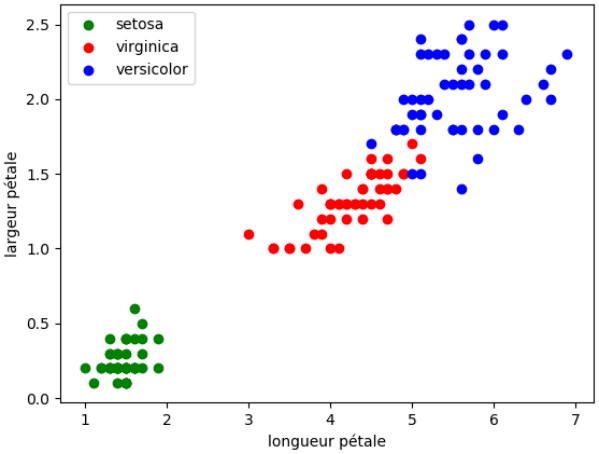
Vous trouverez 50 de ces mesures dans le fichier simplifié iris2D.csv
Extrait du jeu de données "iris" :

	A	B	C
1	longueur p.	largeur p.	espece
2	1.4	0.2	setosa
3	1.4	0.2	setosa
4	1.3	0.2	setosa
5	1.5	0.2	setosa

- la longueur des pétales
- la largeur des pétales
- l'espèce de l'iris

Ce jeu de données a, aujourd'hui, un intérêt purement pédagogique. En effet, il est exclusivement utilisé par des personnes désirant s'initier aux algorithmes de ***machine learning.***

1. Obtenir une représentation graphique (longueur largeur) des données contenues dans le fichier iris2D .csv
On choisira une couleur différente pour chaque espèce.

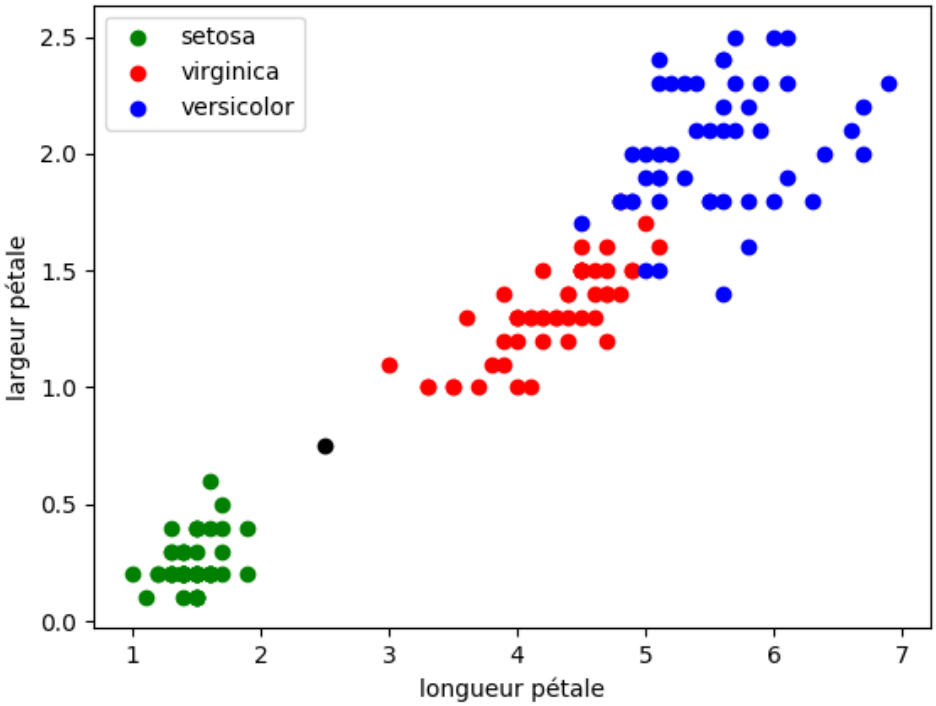


Nous obtenons des "nuages" de points. On remarque ces points sont regroupés par espèces d'iris (sauf pour "iris

virginica" et "iris versicolor", les points ont un peu tendance à se mélanger).

2. Imaginez maintenant qu'au cours d'une promenade vous trouviez un iris, n'étant pas un spécialiste, il ne vous est pas vraiment possible de déterminer l'espèce. En revanche, vous êtes capables de mesurer la longueur et la largeur des pétales de cet iris. Partons du principe qu'un pétale fasse 0,5 cm de large et 2 cm de long.
   Plaçons cette nouvelle donnée sur notre graphique :
   il suffit d'ajouter la ligne
   plt.scatter(2.0, 0.5, color='k')
   et le nouveau point va apparaitre en noir (color='k')) .
   Conclure sur la prédiction dans ce cas simple.

3. Il est possible de rencontrer des cas plus difficiles, par exemple : largeur du pétale = 0,75 cm ; longueur du pétale = 2,5 cm . Faire apparaitre le point.



4. Déterminer graphiquement à l'aide d'un compas les 3 plus proches voisins.

Dans ce genre de cas, il peut être intéressant d'utiliser l'algorithme des "k plus proches voisins" :

- on calcule la « distance » entre notre point (largeur du pétale = 0,75 cm ; longueur du pétale = 2,5 cm) et chaque point issu du jeu de données "iris"
- on sélectionne uniquement les k distances les plus petites (les k plus proches voisins)
- parmi les k plus proches voisins, on détermine quelle est l'espèce majoritaire. On associe à notre "iris mystère" cette "espèce majoritaire parmi les k plus proches voisins"

5. Ecrire un programme en python qui permettent de déterminer l'espèce de l'iris pour k=1,3,5 et 7 (on affichera aussi les valeurs de chacun des voisins)

6. Modifier le programme précédent pour déterminer les plus proches voisins en prenant en compte les 4 mesures pour le fichier iris.csv